Statistique et analyse des données

Vassili Giakoumakis
Bernard Monjardet

Coefficients d'accord entre deux préordres totaux

Statistique et analyse des données, tome 12, n° 1-2 (1987), p. 46-99
<http://www.numdam.org/item?id=SAD_1987__12_1-2_46_0>

Statistique et Analyse des Données
1987 - Vol.12 n° 1 et 2 - pp. 46-99

COEFFICIENTS D'ACCORD ENTRE DEUX PREORDRES TOTAUX

Vassili GIAKOUMAKIS, Bernard MONJARDET

Université Paris 5 et Centre Analyse et Mathématiques Sociales
54 bd. Raspail - 75270 PARIS CEDEX 06

Résumé : *On présente seize coefficients d'accord entre préordres totaux dont la plupart ont été proposés dans la littérature à propos de problèmes variés (analyse de préférences, mesure d'association de deux variables ordinales, comparaison de préordonnances, mesures d'ajustement ...) Ces coefficients sont classés selon deux critères, dichotomiques et indépendants : d'une part un tel coefficient est obtenu en normalisant soit une dissimilarité (ou une similarité) entre préordres totaux, soit un produit scalaire de vecteurs codant ces préordres totaux; d'autre part il généralise soit le tau de Kendall, soit le rho de Spearman entre ordres totaux. Chaque coefficient est étudié individuellement, notamment pour déterminer les couples de préordres totaux donnant ses valeurs extrêmes* (± 1) *, et si possible situé parmi les coefficients construits de manière analogue. En conclusion, on indique comment les résultats obtenus peuvent aider dans le choix du coefficient adapté à un problème particulier et on évoque des directions de recherche dans cet esprit.*

Abstract: *We present sixteen agreement coefficients between weak orders (transitive and complete binary relations) defined on an arbitrary set. Most of them have been defined by various authors in connection with problems like preferences analysis, correlation or association measures between ordinal variables, comparison of dissimilarities, ordinal measures of fit and so on. We classify these coefficients according to two criteria : first*

Manuscrit reçu le 14.10.86, révisé le 29.10.87

the coefficient is a generalization of either the Kendall tau or the Spearman rho (for linear orders); secondly it is obtained by normalizing either a dissimilarity measure between weak orders or a scalar product between vectors coding weak orders. For each coefficient we characterize the pair of weak orders giving the extremal values ± 1 *. A fundamental question is : what coefficient choose in a specific situation ? We give some beginnings of an answer to this question and we point out further work and research directions about it.*

Mots clés : *Préordre total, coefficient d'accord, dissimilarité.*

Indices de classification STMA : *06-020*

1 - INTRODUCTION

Les préordres totaux sont des données relationnelles rencontrées fréquemment. On peut en effet les obtenir, soit directement, lorsque par exemple la préférence exprimée par un individu entre divers objets possibles de choix est transitive et totale, soit indirectement lorqu'à une fonction numérique quelconque $f : X \to \mathbb{R}$, on associe la relation $xP_f y$ si et seulement si $f(x) \leq f(y)$. Cette dernière procédure est utilisée en statistique chaque fois qu'on estime pertinent de se limiter à la seule information "ordinale" apportée par les données (cf. sur ce point Sibson 1972). C'est par exemple le cas lorsqu'à un tableau de dissimilarités - ou de similarités - entre objets, on associe sa préordonnance, i.e. le préordre total défini sur les paires d'objets par $\{x,y\}P_d\{z,t\}$ si et seulement si $d(x,y) \leq d(z,t)$ (cf. notamment Shepard, 1962, Kruskal, 1964, De la Vega, 1967, Lerman, 1970, Benzécri,1973). De même plusieurs auteurs (cf. notamment Mirkin, 1979) ont utilisé la transformation qui à un tableau croisé de n objets et p variables numériques associe les p préordres totaux définis sur ces objets par ces variables.

Le problème de définir un coefficient d'accord entre deux préordres totaux a aussi été souvent posé, et dans des contextes variés. Il peut s'agir, par exemple, de mesurer l'accord entre les préférences de deux sujets différents, ou du même sujet à deux moments différents. Il peut aussi s'agir de

disposer d'une mesure de proximité qu'on cherchera ensuite à optimiser dans une procédure d'ajustement d'un modèle aux données; c'est ainsi qu'on peut chercher la préordonnance "ultramétrique" la plus proche d'une préordonnance donnée (cf. Schader, 1980) ou - cas particulier du précédent - la partition la plus proche de cette préordonnance (cf. en particulier, De la Vega, 1967, Lerman, 1970, Benzecri, 1973, Chah, 1984, Chandon et Boctor, 1985). La multiplicité des situations où l'on a besoin d'un coefficient d'accord entre deux préordres totaux explique sans nul doute la variété de tels coefficients rencontrée dans la littérature, variété où il est bien difficile de s'orienter pour trouver le coefficient le mieux adapté à son problème. Disons tout de suite en effet que pour nous il n'y a pas de coefficient "universellement" meilleur, et que le choix du bon coefficient dépend essentiellement du contexte particulier où sont définis les préordres totaux considérés et de l'usage à quoi on le destine. Encore faut-il pour faire le bon choix que l'on connaisse les coefficients disponibles et leurs propriétés. Dans ce texte nous présentons seize tels coefficients et certaines de leurs propriétés; la plupart des coefficients que nous étudions ont déjà été définis dans la littérature, mais nous en avons introduit quelques autres tout aussi naturels que les précédents. Ces seize coefficients peuvent être classés selon les deux critères dichotomiques suivants :

1. ils généralisent soit l'un soit l'autre des deux coefficients d'accord les plus classiques entre ordres totaux : le tau de Kendall et le rho de Spearman.
2. Ils sont obtenus en normalisant soit une dissimilarité (ou une similarité) entre préordres totaux, soit un produit scalaire entre vecteurs codant les préordres totaux.

Nous présentons maintenant le contenu des différents paragraphes de ce texte. Le paragraphe 2 sert d'abord à introduire nos notations sur les préordres totaux et à rappeler quelques notions fondamentales, notamment la décomposition de l'ensemble X^2 des couples de X associée à la donnée d'un préordre total P sur X : $X^2 = O + I + O^r$ (O partie asymétrique de P, I partie symétrique). On présente ensuite les codages qui à un préordre total défini sur un ensemble à n éléments associe un vecteur de l'espace vectoriel $\mathbb{R}^{n^2}$ (codage α, β et γ) ou $\mathbb{R}^n$ (codage du rang

moyen). Puis on définit la partition fondamentale de l'ensemble X^2 , associée à la donnée de deux préordres totaux sur X ; les effectifs des neuf classes de cette partition définissent cinq paramètres fondamentaux. On explicite ensuite les paires particulières de préordres totaux obtenus lorsqu'un ou plusieurs de ces paramètres sont nuls. Le paragraphe 3 est consacré à l'étude de coefficients d'accord obtenus par la normalisation d'une dissimilarite entre préordres totaux la faisant varier de -1 (lorsque la dissimilarité entre les deux préordres est maximum) à $+1$ (lorsque cette dissimilarité est nulle). Les dissimilarités considérées peuvent être la distance de la norme L_1 , ou le carré de la distance euclidienne dans les espaces vectoriels de codages (et en particulier la classique distance de la différence symétrique); mais d'autres possibilités sont aussi examinées. On obtient ainsi les sept coefficients τ_1 à τ_7 généralisant tau , dont l'étude individuelle, puis comparée est faite dans la section 3.3, notamment pour montrer comment ils pondèrent différemment les classes de la partition fondamentale. La section 3.4 est consacrée à un coefficient ρ_1 généralisant rho . Dans la section 4, on étudie des coefficients d'accord variant toujours entre -1 et $+1$ mais obtenus par la normalisation du produit scalaire des deux vecteurs codant les préordres totaux. On peut en particulier prendre le coefficient de corrélation de ces deux vecteurs, ce qui conduit notamment aux coefficients τ_b et ρ_b de Kendall. Mais on considère aussi d'autres possibilités classiques, permettant de définir cinq coefficients (τ_a , e , d_a , d_b et γ) généralisant tau , et plus généralement les "coefficients de monotonicité" de Guttman, ou les coefficients ρ_a et τ_γ .

En conclusion on montre comment les résultats obtenus peuvent aider dans le choix d'un coefficient d'accord entre deux préordres totaux et on indique des travaux ultérieurs ou des directions de recherche permettant d'avancer dans cette direction.

Précisons en terminant cette introduction que les seize coefficients considérés ici sont loin d'être les seuls possibles. On aura d'ailleurs l'occasion au cours du texte d'en citer quelques autres dont certains sont étudiés dans Giakoumakis (1985). D'autre part les normalisations que nous avons considérées ne font pas intervenir la distribution statistique de la mesure

"brute" d'accord choisie, alors qu'on peut normaliser à partir de cette dernière (cf. notamment Lerman, 1973, 1981, Le Calve, 1976). Par contre il n'est pas rare de trouver le même coefficient sous des formes et des appellations variées chez différents auteurs. Notre étude permettra nous l'espérons de remédier à cette situation.

2 - NOTIONS DE BASE

2.1 - Rappels et notations

Dans tout ce texte, X désigne un ensemble fini, à n éléments. Une relation binaire R sur X est un ensemble de couples de X, i.e. une partie de X^2. On note indifféremment $(x,y) \in R$, ou xRy. On note R^c la relation complémentaire de R : $(x,y) \in R^c$ si et seulement si $(x,y) \notin R$; on note R^r la relation réciproque de R : $(x,y) \in R^r$ si et seulement si $(y,x) \in R$; enfin on pose $R^d = (R^r)^c$ et on appelle cette relation la duale de R. $|R|$ désigne le nombre de couples de la relation R.

Un *préordre total* P sur X est une relation binaire sur X, réflexive (xRx, pour tout x), transitive (xRy et yRz impliquent xRz) et totale (xR^cy implique yRx). Si, de plus P est antisymétrique (xRy et yRx impliquent $x = y$) P est un *ordre total*.

Une *partition ordonnée* de l'ensemble X est la donnée d'une partition $\pi = \{C_1, C_2, \ldots, C_t\}$ de X et d'un ordre total noté $\underset{\pi}{\leq}$ sur les classes de cette partition. On la note $(\pi, \underset{\pi}{\leq})$, ou en supposant que les classes sont numérotées suivant l'ordre $\underset{\pi}{\leq}$, $C_1 < C_2 \ \ldots < C_t$.

Les propriétés classiques des préordres totaux sont bien connues (cf. par exemple, Barbut et Monjardet, 1970). Nous introduisons les notations utilisées dans ce texte en rappelant quelques faits fondamentaux.

- Tout préordre total P est égal à $I + O$, où
  $I = \{(x,y) \in X^2$ tel que $(x,y) \in P$ et $(y,x) \in P\}$ (partie symétrique de P)
  $O = \{(x,y) \in X^2$ tel que $(x,y) \in P$ et $(y,x) \notin P\}$ (partie asymétrique de P)

+ est la notation pour la réunion de deux ensembles disjoints.

- Un préordre total $P = I + O$ induit une partition de l'ensemble des couples de X : $X^2 = O + I + O^r$.

- La partie symétrique I de P est une relation d'équivalence (i.e. une relation réflexive, symétrique, et transitive) à laquelle est associée une partition π . Par définition, les classes du préordre P , sont les classes de cette partition π (i.e. les classes de l'équivalence I). On aura parfois besoin de considérer la relation I sans ses couples (x,x) de réflexivité; on la notera alors $\overset{o}{I}$. De même, $\overset{o}{P}$ désignera le préordre P sans ses couples de réflexivité.

- Le préordre total P induit un ordre total sur les classes, donc une partition ordonnée de X . Inversement à toute partition ordonnée de X correspond un préordre total sur X (cette correspondance étant bijective). Ce fait justifie qu'un préordre total P soit souvent noté par la partition ordonnée associée, i.e. par $C_1 < C_2 \ldots < C_t$ (où les C_i sont les classes de P) . En particulier, si toutes les classes C_i sont à un seul élément, i.e. si P est un ordre total, on le notera $x_1 < x_2 \ldots < x_n$, ou plus simplement $x_1 x_2 \ldots x_n$ (notation par un "mot" ou une "permutation" de X) .

- La partie asymétrique O du préordre total P est un ordre strict (i.e. une relation transitive et asymétrique). En fait, un tel ordre n'est pas quelconque; il a la structure d'un "ordre fort", type d'ordre partiel qu'on peut caractériser de plusieurs façons (cf. par exemple, Leclerc et Monjardet, 1973). On notera que O est la relation duale de P : xOy si et seulement si $yP^c x$.

- Finalement, nous notons X^2 le préordre total contenant tous les couples de X , i.e. le préordre P pour lequel $P = I = X^2$ et $O = \emptyset$.

2.2 - Codages des préordres totaux

On note $\mathcal{P}$ l'ensemble de tous les préordres totaux définis sur X (pour le dénombrement et la structure de $\mathcal{P}$ voir Barbut, Monjardet, 1970 et

Giakoumakis, 1985). On appelle *codage* des préordres totaux une application c qui à tout préordre total fait correspondre un vecteur $\vec{c}$ de l'espace euclidien $\mathbb{R}^p$:

$$\mathcal{P} \longrightarrow \mathbb{R}^p$$
$$P \longrightarrow c(P) = \vec{c}$$

Dans cet article, on utilisera deux types de codage. Le premier code un préordre total en codant tous les couples de X^2 (i.e. en faisant correspondre à tout $(x,y) \in X^2$ un nombre réel); il est donc dans $\mathbb{R}^{n^2}$. Le deuxième code un préordre total en codant tous les éléments de X ; il est donc dans $\mathbb{R}^n$.

Dans tous les cas on se donne un ordre total arbitraire, mais fixe,sur l'ensemble des éléments à coder, i.e. sur X^2 ou sur X .

Codage des préordres totaux dans $\mathbb{R}^{n^2}$

Le codage α (codage caractéristique)

L'image notée $\vec{\alpha}$ d'un préordre total P par ce codage est définie de la façon suivante : la coordonnée indicée par le couple $(x,y) \in X^2$ est notée $\alpha(x,y)$ et est donnée par :

$\alpha(x,y) = +1 \Leftrightarrow (x,y) \in P$

$\alpha(x,y) = 0 \Leftrightarrow (x,y) \notin P$

On a : $\|\vec{\alpha}\| = (|P|)^{1/2}$ et $\sum_{(x,y)} \alpha(x,y) = |P|$

Le codage β

Le vecteur $\vec{\beta}$ associé à P par le codage β est défini par :

$\beta(x,y) = +1 \Leftrightarrow (x,y) \in P$ et $(y,x) \notin P \Leftrightarrow (x,y) \in O$

$\beta(x,y) = 0 \Leftrightarrow (x,y) \in P$ et $(y,x) \in P \Leftrightarrow (x,y) \in I$

$\beta(x,y) = -1 \Leftrightarrow (x,y) \notin P$ et $(y,x) \in P \Leftrightarrow (x,y) \in O^r$

(O^r est la relation réciproque de O)

On a : $\sum_{(x,y)} \beta(x,y) = 0$, donc $\vec{\beta}$ est un vecteur centré et

$\|\vec{\beta}\| = n^2 - (|I|)^{1/2} = (2|O|)^{1/2}$

Le codage γ (codage uniforme)

Le vecteur $\vec{\gamma}$ associé à P par ce codage est défini par :

$\gamma(x,y) = +1 \Leftrightarrow (x,y) \in P$ et $x \neq y$

$\gamma(x,y) = -1 \Leftrightarrow (x,y) \notin P$ et $x \neq y$

$\gamma(x,y) = 0 \Leftrightarrow x = y$

On a : $\|\vec{\gamma}\| = (n(n-1))^{1/2}$ et $\sum_{(x,y)} \gamma(x,y) = |\overset{o}{I}|$

Remarques :

1. Le codage β est relié au codage caractéristique α par la formule $\beta(x,y) = \alpha(x,y) - \alpha(y,x)$. Par contre, il n'existe aucune relation linéaire ou affine entre ces trois codages, sauf si P est un ordre total (auquel cas $\vec{\beta} = \vec{\gamma}$) , ou si $P = X^2$ (auquel cas $\vec{\alpha} = \vec{1}+\vec{\beta}$, $\vec{1}$ désignant le vecteur dont toutes les coordonnées sont égales à 1).

2. Le codage α correspond au classique vecteur caractéristique d'une partie d'un ensemble, ou encore à la matrice d'incidence usuelle d'une relation binaire; une variante de α consisterait à coder o les couples (x,x) . Le codage β , utilisé notamment par Kendall (1970) est le plus classique.

   Une variante affine de β est le codage $\frac{\vec{\beta} + \vec{1}}{2}$ qui code les couples de O,I et O^r par respectivement $+1$, $+\frac{1}{2}$ et 0 . Le codage γ a l'intérêt de coder les préordres totaux dans une hypersphère de $\mathbb{R}^{n^2}$.

3. Il existe bien sûr d'autres codages possibles dans $\mathbb{R}^{n^2}$. Citons par exemple le codage "caractéristique" de l'ordre O associé au préordre P (1 si $(x,y) \in O$ et 0 sinon). Toutefois, compte tenu de la dualité entre O et P , certains des coefficients d'accord définis à partir de ce codage pourront être les mêmes que ceux définis à partir du codage β . (cf. *Synthèse* dans 3.3).

Codages des préordres totaux dans $\mathbb{R}^n$

On a d'abord besoin de la notion de rang d'un élément dans un préordre total P . Soit x élément de X ; le *rang* dans P de x , noté $r^-(x)$ est défini par

$$r^-(x) = |\{y \in X : yPx\}|$$

Si $P = C_1 < C_2 \ldots < C_t$ et si on pose $n_\ell = |C_\ell|$, pour $\ell = 1,\ldots,t$ on a clairement, pour $x \in C_k$, $r^-(x) = \sum_{\ell=1}^{k} n_\ell$

Pour définir les codages ci-dessous, on se donne un ordre total arbitraire (mais fixé) sur X, et on le note $x_1x_2\ldots x_i\ldots x_n$.

Le codage "rang moyen"

L'image $r(P)$ d'un préordre total $P = C_1 < C_2 \ldots < C_t$ est un vecteur de $\mathbb{R}^n$, noté $\vec{r} = (r_1,\ldots r_i,\ldots r_n)$ et défini de la manière suivante :

soit C_k la classe de P à laquelle appartient x_i,

$$r_i = r(x_i) = \frac{\sum_{j=1}^{n_k} (r^-(x_i)-j+1)}{n_k}$$

$r(x_i)$ s'appelle le *rang moyen* de x_i : c'est en fait la moyenne arithmétique des rangs qu'auraient les éléments de la classe C_k s'ils étaient totalement ordonnés (dans le préordre P). On a les propriétés suivantes pour ce codage :

- $r_i = r^-(x_i) - \frac{n_k-1}{2} = \sum_{\ell<k} n_\ell + \frac{n_k+1}{2} \quad (x_i \in C_k)$

  En particulier, $r_i = r^-(x_i)$, pour tout i, si et seulement si P est un ordre total.

- $\sum_{i=1}^{n} r_i = \frac{n(n+1)}{2}$

- La somme des carrés des rangs moyens des éléments de la classe C_k est plus petite que la somme des carrés des rangs qu'auraient ces éléments s'ils étaient totalement ordonnés, la différence des deux sommes étant $\frac{1}{12}(n_k^3 - n_k)$.

  On en déduit les expressions suivantes de la norme euclidienne de $\vec{r}$:

- $\|\vec{r}\| = \left[\sum_{k=1}^{t} n_k \left(\sum_{\ell<k} n_\ell + \frac{n_k+1}{2}\right)^2\right]^{\frac{1}{2}} = \left[\frac{2n(n+1)(2n+1) - \sum_{k=1}^{t}(n_k^3-n_k)}{12}\right]^{\frac{1}{2}}$

Le codage score

Le *score* d'un élément x dans un préordre total $P = C_1 < C_2 \ldots < C_t$, noté $s(x)$, est défini par

$$s(x) = |\{y \in X : xPy\}| - |\{y \in X : yPx\}|$$

L'image $s(P)$ d'un préordre total P par le codage score est le vecteur $\vec{s} : (s_1,\ldots s_i,\ldots s_n)$ de $\mathbb{R}^n$, défini par

$$s_i = s(x_i) \text{ , pour tout } i \text{ .}$$

On a clairement, pour $x_i \in C_k$

$$s_i = n + n_k - 2r^-(x_i) = n + 1 - 2r_i \text{ .}$$

D'où $\vec{s} = (n\vec{+}1) - 2\vec{r}$ est un vecteur centré, et on obtient les expressions suivantes de sa norme :

$$\|\vec{s}\| = \left[\sum_{k=1}^{t} n_k \left(2 \sum_{\ell<k} n_\ell + n_k + 1\right)^2 - n(n+1)^2\right]^{\frac{1}{2}} = \left[\frac{n^3 - n - \sum_{k=1}^{t} (n_k^3 - n_k)}{3}\right]^{\frac{1}{2}}$$

Dans la suite de ce travail, nous n'utilisons pas le codage score. En effet, compte tenu de la liaison affine entre $\vec{s}$ et $\vec{r}$, on montre aisément que les coefficients d'accord définis ultérieurement, et basés soit sur $\vec{r}$, soit sur $\vec{s}$, sont identiques.

Remarques

1. Le codage rang moyen remonte au moins à Kendall (1970). Le codage score est, par exemple, utilisé dans Cailliez et Pages (1976) ou Lemaire (1977).
2. On peut dire qu'un codage d'un préordre total $P = I + O$ dans $\mathbb{R}^n$ est admissible si et seulement si pour xIy on a $c(x) = c(y)$ et pour xOy on a soit (toujours) $c(x) < c(y)$, soit (toujours) $c(x) > c(y)$. Il est clair que les deux codages précédents sont admissibles. Dans certaines méthodes d'analyse des données, notamment en régression, on recherche un tel codage qui soit optimal par rapport à un certain critère.
3. Le codage rang moyen est relié au codage β par la formule

$$r(x) = \frac{n+1 - \sum_{y \in X} \beta(x,y)}{2}$$

Exemples

1. Codages dans $\mathbb{R}^{n^2}$

Soient $X = \{a,b,c\}$, $P = a < bc$ et $P' = ac < b$

	(a,a)	(a,b)	(a,c)	(b,b)	(b,a)	(b,c)	(c,c)	(c,b)	(c,a)
codage α									
a < bc	1	1	1	1	0	1	1	1	0
ac < b	1	1	1	1	0	0	1	1	1
codage γ									
a < bc	0	1	1	0	-1	1	0	1	-1
ac < b	0	1	1	0	-1	-1	0	1	1
codage β									
a < bc	0	1	1	0	-1	0	0	0	-1
ac < b	0	1	0	0	-1	-1	0	1	0

2. Codages dans $\mathbb{R}^n$

Soient $X = \{a,b,c,d,e,f\}$, $P = a < bc < df < e$ et $P' = f < de < b < a < c$.

	a	b	c	d	e	f
codage rang moyen						
a < bc < df < e	1	2,5	2,5	4,5	6	4,5
f < de < b < a < c	5	4	6	2,5	2,5	1
codage score						
a < bc < df < e	5	2	2	-2	-5	-2
f < de < b < a < c	-3	-1	-5	2	2	5

N.B. Pour tous les coefficients d'accord définis dans ce texte, nous donnerons comme illustration la valeur du coefficient pour les deux préordres P et P' ci-dessus. Nous dirons que ces deux préordres constituent l'exemple de référence.

2.3 - Paires de préordres totaux

Partitions et paramètres fondamentaux associés à une paire de préordres totaux

Soit $P = O + I$, $P' = O' + I$ deux préordres totaux définis sur X. On appelle partition (fondamentale) associée à la paire (P,P'), la partition des couples d'éléments de X induite par le croisement des deux par-

titions associées à P et P' . Elle est représentée par le tableau 1 ci-dessous.

	O'	I'	O'^r	
O	$O \cap O'$	$O \cap I'$	$O \cap O'^r$	P
I	$I \cap O'$	$I \cap I'$	$I \cap O'^r$	
O^r	$O^r \cap O'$	$O^r \cap I'$	$O^r \cap O'^r$	P^r
	P'		P'^r	

Tableau 1

On obtient donc, en général, une partition de X^2 en neuf classes. Il est commode de donner des noms aux couples d'éléments de X suivant la classe à laquelle ils appartiennent.

Etant donné la paire (P,P') de préordres totaux, nous dirons que :

(x,y) est un *accord strict* ssi $(x,y) \in O \cap O'$
(x,y) est un *accord large* ssi $(x,y) \in I \cap I'$
(x,y) est un *semi-accord* ssi $(x,y) \in O \cap I' + I \cap O'$
(x,y) est un *semi-désaccord* ssi $(x,y) \in O^r \cap I' + I \cap O'^r$
(x,y) est un *désaccord strict* ssi $(x,y) \in O \cap O'^r + O^r \cap O'$
(x,y) est un *co-accord* ssi $(x,y) \in O^r \cap O'^r$

Avec ces notations, on voit que l'intersection $P \cap P'$ des deux préordres totaux est l'ensemble des couples en accord (strict, large, ou semi) :

$$P \cap P' = O \cap O' + I \cap I' + I \cap O' + I' \cap O$$

De même la différence symétrique $P \Delta P'$ (i.e. $P \cap P'^c + P^c \cap P'$) est l'ensemble des couples en désaccord (strict ou semi) :

$$P \Delta P' = O \cap O'^r + O^r \cap O' + I \cap O'^r + O^r \cap I'$$

Enfin, les couples en co-accord sont les couples n'appartenant ni à P , ni à P' :

$$P^c \cap P'^c = O^r \cap O'^r \ .$$

De nombreux coefficients d'accord entre deux préordres totaux sont calculés à partir des effectifs de certaines des classes précédentes. Ces ef-

fectifs jouent un rôle de paramètres fondamentaux et nous allons leur donner des notations spécifiques. Auparavant, remarquons que certains des effectifs précédents sont égaux. En effet, l'application réciproque (définie en 2.1) induit une bijection entre certaines des classes. Précisément, on a

$$(O \cap O')^r = O^r \cap O'^r \qquad (O^r \cap O')^r : O \cap O'^r$$
$$(I \cap O')^r = I \cap O'^r \qquad (O^r \cap I')^r : O \cap I'$$

Compte tenu de ces relations, on a finalement (au plus) cinq paramètres différents que nous notons a, b_1, b_2, c, d . Le tableau 2 ci-dessous explicite ces notations :

P \ P'	O'	I'	O'^r
O	a	b_2	d
I	b_1	c	b_1
O^r	d	b_2	a

Tableau 2

Ainsi :

- a est le nombre d'accords stricts (ou de co-accords)
- c est le nombre d'accords larges
- 2d est le nombre de désaccords stricts
- b_1+b_2 , noté b , est le nombre de semi-accords (ou de semi-désaccords) .

D'autre part, nous posons (rappelons que $n = |X|$)

$$c^* = \frac{c-n}{2} = \frac{|I^o \cap I'^o|}{2}$$

Enfin, on a les relations suivantes entre les paramètres fondamentaux et le nombre n d'éléments de X :

$$n^2 = 2(a + b + d) + c$$
$$n(n-1) = 2(a + b + c^* + d)$$

Noter aussi qu'on a $|P \cap P'| = a + b + c$, $|P \Delta P'| = b + 2d$.

Remarque

L'analyse précédente est en termes de couples d'éléments de X . On peut faire une analyse en termes de paires. Il suffit pour cela de considérer la

partition en quatre classes des paires de X, obtenue en prenant les réunions d'une des catégories précédentes avec la catégorie réciproque (et en excluant les couples (x,x)). On pourra alors dire que pour une paire (x,y), il y a soit accord strict ($O \cap O'$ et $O^r \cap O'^r$), soit accord large ($\overset{o}{I} \cap \overset{o}{I'}$), soit désaccord strict ($O \cap O'^r + O^r \cap O'$) soit semi-accord ou semi-désaccord ($O \cap I' + I \cap O' + O^r \cap I' + I \cap O'^r$). Dans ce cas, il n'y a plus que quatre paramètres fondamentaux correspondant aux effectifs de ces quatre classes : a, c^*, d et b. Nous verrons que la plupart des coefficients d'accord définis dans ce texte s'expriment en fonction de ces quatre nombres.

Pour calculer les paramètres fondamentaux on peut utiliser des formules les donnant en fonction des effectifs des classes de l'équivalence $I \cap I'$. Nous allons donner ci-dessous ces formules. Soient deux préordres totaux

$$P = O + I = C_1 < C_2 \ldots < C_i \ldots < C_t$$
$$P' = O' + I' = C'_1 < C'_2 \ldots < C'_j \ldots < C'_q$$

On leur associe le tableau croisé des effectifs des classes $C_i \cap C'_j$:

	C'_1	C'_j	C'_q
C_1			
C_i		n_{ij}	
C_t			

On pose

$$n_{ij} = |C_i \cap C'_j|$$

$$n_i = \sum_{j=1}^{q} n_{ij} = |C_i| \qquad n'_j = \sum_{i=1}^{t} n_{ij} = |C'_j|$$

On a

$$|I| = \sum_i n_i^2 \qquad |I'| = \sum_j n_j'^2$$

$$|O| = |O^r| = \frac{1}{2}(n^2 - \sum_i n_i^2) \qquad |O'| = |O'^r| = \frac{1}{2}(n^2 - \sum_j n_j^2)$$

On obtient alors les formules suivantes :

$$a = \sum_{i,j} [n_{ij} (\sum_{\substack{k>i \\ \ell>j}} n_{k\ell})] \qquad d = \sum_{i,j} [n_{ij} (\sum_{\substack{k<i \\ \ell>j}} n_{k\ell})]$$

$$b_1 = \frac{1}{2} (\sum_i n_i^2 - \sum_{i,j} n_{ij}^2) \qquad b_2 = \frac{1}{2} (\sum_j n_j^2 - \sum n_{ij}^2)$$

$$b = \frac{1}{2} (\sum_i n_i^2 + \sum_j n_j^2) - \sum_{i,j} n^2{}_{ij} \qquad c = \sum_{i,j} n_{ij}^2 \qquad c^* = \sum_{i,j} \binom{n_{ij}}{2}$$

D'où

$$|P \cap P'| = \sum_{i,j} [n_{ij} (\sum_{\substack{k>i \\ \ell>j}} n_{k\ell})] + \frac{1}{2} (\sum_i n_i^2 + \sum_j n_j^2)$$

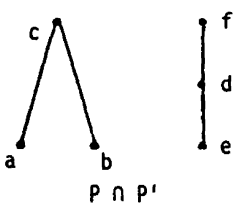
$$|P \Delta P'| = n^2 - 2 \sum_{i,j} [n_{ij} (\sum_{\substack{k>i \\ \ell>j}} n_{k\ell})] - \frac{1}{2} (\sum_i n_i^2 + \sum_j n_j^2)$$

Exemple :

Dans l'exemple de référence où $P = a < bc < dc < f$ et $P' = e < df < b < a < c$, on obtient les tableaux suivants :

	e	df	b	a	c
a	0	0	0	1	0
bc	0	0	1	0	1
de	1	1	0	0	0
f	0	1	0	0	0

	0'	I'	$0'^r$
0	2	1	10
I	2	6	2
0^r	10	1	2



Nous avons aussi représenté $P \cap P'$ (qui dans ce cas est un ordre partiel). On a $|0| = 13$, $|I| = 10$, $|0'| = 14$, $|I'| = 8$, $|P \cap P'| = 11$, $|P \Delta P'| = 23$.

Paires particulières de préordres totaux

Nous dirons que (P,P') est une paire particulière de préordres totaux si l'une ou plusieurs des classes de la partition fondamentale associée sont

vides, ou équivalemment si certains des paramètres fondamentaux sont nuls. Nous donnons ci-dessous la liste de telles paires. Pour chaque type de paire, on indique les paramètres qui sont nuls, et on donne une caractérisation "combinatoire" de ce type de paire. Par exemple, $d = 0$ signifie qu'on a une paire P,P' avec le paramètre $d(P,P')$ nul, ce qui est équivalent au fait que pour ces deux préordres totaux, leur intersection $P \cap P'$ est un préordre total.

- $a = 0 \Leftrightarrow 0 \cup 0'^r \subseteq P \cap P'^r \Leftrightarrow P \cap P'^r$ préordre total
- $d = 0 \Leftrightarrow 0 \cup 0' \subseteq P \cap P' \Leftrightarrow P \cap P'$ préordre total
- $b_1 = 0 \Leftrightarrow I \subseteq I'$; $b_2 = 0 \Leftrightarrow I \supseteq I'$
  $b = 0 \Leftrightarrow I = I'$
- $c^* = 0 \Leftrightarrow c = n \Leftrightarrow \overset{o}{I} \cap \overset{o}{I'} = \emptyset \Leftrightarrow P \cap P'$ ordre
- $a = b_1 = 0 \Leftrightarrow P \subseteq P'^r$ $a = b_2 = 0 \Leftrightarrow P \supseteq P'^r$
  $a = b = 0 \Leftrightarrow P = P'^r$
- $d = b_1 = 0 \Leftrightarrow P \subseteq P'$ $d = b_2 = 0 \Leftrightarrow P \supseteq P'$
  $d = b = 0 \Leftrightarrow P = P'$
- $a = c^* = 0 \Leftrightarrow P \cap P'^r$ ordre total $d = c^* = 0 \Leftrightarrow P \cap P'$ ordre total
- $b_1 = c^* = 0 \Leftrightarrow P$ ordre total $b_2 = c^* = 0 \Leftrightarrow P'$ ordre total
  $b = c^* = 0 \Leftrightarrow P$ et P' ordres totaux
- $a = d = 0 \Leftrightarrow P = X^2$ ou $P' = X^2$
- $a = b = c^* = 0 \Leftrightarrow P = P'^r$ ordre total ;
  $b = c^* = d = 0 \Leftrightarrow P = P'$ ordre total
- $a = b = d = 0 \Leftrightarrow P = P' = X^2$
- $a = c^* = d = 0 \Leftrightarrow P = X^2$ et P' ordre total (ou $P' = X^2$ et P ordre total).

La plupart des caractérisations ci-dessus sont évidentes ou faciles à démontrer. Noter que certaines se déduisent immédiatement d'autres en utilisant les égalités :

$$a(P,P') = d(P,P'^r) = a(P^r,P'^r)$$

On pourra comme exercice montrer que l'égalité $a = d = 0$ implique que P ou P' égalent X^2, et préciser la structure des paires de préordres totaux vérifiant $a = c^* = 0$ (ou $d = c^* = 0$).

N.B. : Il résulte des caractérisations ci-dessus, qu'on a $d = 0$ et $a \neq 0$ si et seulement si $P \cap P'$ est total avec P et P' différents de X^2.

De même, $a = c$ et $d \neq 0$, si et seulement si $P \cap P'^r$ total, avec P et $P' \neq X^2$. Ces types de paires seront considérés dans 4.2 au paragraphe *Les coefficients.*

3 - COEFFICIENTS D'ACCORD DEFINIS PAR LA NORMALISATION D'UNE DISSIMILARITE (OU D'UNE SIMILARITE) ENTRE PREORDRES TOTAUX

3.1 - Forme générale des coefficients

Tous ces coefficients sont définis en normalisant une "mesure de désaccord" δ entre deux préordres totaux, i.e. une application qui à tout couple (P,P') de préordres totaux associe un nombre positif ou nul $\delta(P,P')$ et qui vérifie la condition :

(0) $\delta(P,P') = \delta(P',P)$

δ peut vérifier certaines des conditions suivantes :

(1) $\delta(P,P') = 0$ si $P = P'$

(2) $\delta(P,P') = 0$ implique $P = P'$

(3) $\delta(P,P') \leq \delta(P,P'') + \delta(P'',P')$

Si δ vérifie (respectivement - ne vérifie pas) la condition (1), nous dirons que c'est une *dissimilarité* (respectivement - une *pseudo-dissimilarité*). Si δ vérifie les quatre conditions, c'est une *distance*; on remarquera que dans ce cas δ^2 vérifie (0), (1) et (2) mais non nécessairement (3).

On définit un coefficient d'accord κ entre préordres totaux en "normalisant" δ de façon que :

$$-1 \leq \kappa(P,P') \leq +1$$

$$\kappa(P,P') = +1 \Leftrightarrow \delta(P,P') = 0$$

$$\kappa(P,P') = -1 \Leftrightarrow \delta(P,P') = \operatorname*{Max}_{P,P' \in \mathcal{P}} \delta(P,P')$$

Nous notons dans la suite $\operatorname{Max}\delta$, le maximum de la dissimilarité (sur toutes les paires de préordres totaux). D'autre part, nous supposons que $\kappa(P,P')$ est une fonction affine de δ, i.e. est de la forme $a+b\delta(P,P')$.

Un calcul évident donne alors :

$$(1) \qquad \kappa(P,P') = 1 - \frac{2\delta(P,P')}{\mathrm{Max}_{\delta}} \quad .$$

A une dissimilarité δ entre préordres totaux est associée une similarité σ obtenue en posant :

$$\sigma(P,P') = \mathrm{Max}\delta(P,P') - \delta(P,P')$$

Si on remplace δ par σ dans l'expression de κ, on obtient (puisque $\mathrm{Max}\sigma = \mathrm{Max}\delta$) :

$$(2) \qquad \kappa(P,P') = \frac{2\sigma(P,P')}{\mathrm{Max}\sigma} - 1 \quad .$$

Enfin, en posant $\theta(P,P') = \sigma(P,P') - \delta(P,P')$ et puisque $\mathrm{Max}\theta = \mathrm{Max}\sigma = \mathrm{Max}\delta$, on obtient aussi :

$$(3) \qquad \kappa(P,P') = \frac{\theta(P,P')}{\mathrm{Max}\theta} \quad .$$

Ainsi, en posant $K = \mathrm{Max}\delta = \mathrm{Max}\sigma = \mathrm{Max}\theta$ on peut écrire, en abrégé :

$$\kappa = 1 - \frac{2\delta}{K} = \frac{2\sigma}{K} - 1 = \frac{\sigma-\delta}{K} \quad .$$

Autrement dit, on a trois expressions pour le coefficient d'accord obtenu en normalisant soit une dissimilarité δ, soit (dualement) une similarité σ entre préordres totaux. Mais, de toutes façons, pour effectuer cette normalisation, on a besoin de calculer le maximum de δ (ou de σ), ce qui n'est pas toujours évident. Par définition de la normalisation, les paires de préordres totaux pour lesquelles le coefficient d'accord est -1 sont celles pour lesquelles $\delta(P,P')$ est maximum. De même, la valeur $+1$ du coefficient est toujours obtenue si $P = P'$ (sauf pour une pseudo-dissimilarité), mais elle peut être obtenue pour d'autres paires si la dissimilarité δ n'est pas une distance ou le carré d'une distance. On remarque aussi qu'on a $\kappa(P,P')$ nul si et seulement si $\delta(P,P') = \sigma(P,P') = \frac{1}{2}\mathrm{Max}\delta$.

3.2 - Dissimilarités (et similarités) entre préordres totaux

Dans la suite, pour définir un coefficient d'accord, nous partirons d'une dissimilarité entre préordres totaux; lorsque ce sera utile, nous donnerons

l'expression de la similarité associée. Comme nous l'avons déjà dit, les dissimilarités utilisées seront souvent des distances. Une manière simple de définir une distance entre préordres totaux est d'utiliser un codage (injectif) de ces relations dans $\mathbb{R}^p$, puis une distance dans l'espace de codage. Nous considérerons trois distances classiques dans $\mathbb{R}^p$: la distance euclidienne notée δ_E, la distance associée à la norme L_1, notée δ_1, et la distance de Hamming (i.e. le nombre de coordonnées différentes), notée δ_H. Dans le cas de la distance euclidienne, il sera plus commode de prendre comme dissimilarité, non cette distance elle-même, mais son carré.

Au paragraphe *Codage des préordres totaux dans* $\mathbb{R}^{n^2}$ de 2.1, nous avions défini trois codages α, β et γ des préordres totaux dans $\mathbb{R}^{n^2}$. Avec les trois distances définies ci-dessus, nous obtenons donc a priori, neuf dissimilarités. En fait, celles-ci se ramènent à trois, car sept d'entre elles sont égales (éventuellement à une transformation linéaire près) à une distance "ensembliste" classique, celle de la "différence symétrique". Nous exposons maintenant ce résultat (Proposition 2, ci-dessous) en commençant par donner différentes expressions de cette distance (Proposition 1). Rappelons que la *distance de la différence symétrique* entre deux préordres totaux, notée δ_Δ, est le cardinal de cette différence :

$$\delta_\Delta(P,P') = |P \Delta P'| = |P \cap P'^C| + |P' \cap P^C| .$$

Proposition 1.

Soient $P = O + I$, $P' = O' + I'$ deux préordres totaux sur X. On a :

(1) $\quad \delta_\Delta(P,P') = 2|O \cap O'| + |I \cap O'^r| + |O^r \cap I'|$

(2) $\quad \delta_\Delta(P,P') = n^2 - 2|O \cap O'| - |I \cap O'| - |I' \cap O| - |I \cap I'|$

(3) $\quad \delta_\Delta(P,P') = \dfrac{\delta_\Delta(I,I')}{2} + 2|O \cap O'^r|$

(4) $\quad \delta_\Delta(P,P') = \delta_\Delta(O,O') = |O| + |O'| - 2|O \cap O'|$

La démonstration de ces formules repose sur les propriétés de la partition fondamentale associée à P et P' (cf. 2.3 - *Partition et paramètres fondamentaux associés à une paire de préordre totaux*). L'expression (1) s'obtient à partir de la définition en remarquant que $P'^C = O'^r$, $P^C = O^r$ et $|O^r \cap O'| = |O \cap O'^r|$. Puisque $P \Delta P' = X^2 - (P \cap P') - (P^C \cap P'^C) =$

$X^2 - (O \cap O') - (I \cap O') - (I' \cap O) - (I \cap I') - (O^r \cap O'^r)$, on obtient (2). On a $\delta_\Delta(I,I') = |I \cap I'^c| + |I^c \cap I'| = |I \cap O'| + |I \cap O'^r| + |I' \cap O| + |I' \cap O^r| = 2(|I \cap O'^r| + |I' \cap O^r|)$, ce qui comparé à (1) donne (3). Enfin (4) s'obtient en remarquant que $O = P^d$, $O' = P'^d$, et que la distance δ_Δ est invariante par dualité.

Dans la formule (3), on peut exprimer la distance de la différence symétrique entre les deux partitions P et P' au moyen des cardinaux de leurs classes, et on obtient ainsi une expression pour δ_Δ équivalente à une formule donnée par Mirkin (1979).

On va maintenant exprimer cette distance de la différence symétrique au moyen des codages α , β et γ (2.2). Nous laissons au lecteur la démonstration du résultat suivant :

Proposition 2.
Soient P et P' deux préordres totaux et $\vec{\alpha}$, $\vec{\alpha}'$, $\vec{\beta}$, $\vec{\beta}'$, $\vec{\gamma}$, $\vec{\gamma}'$, les vecteurs codant ces préordres dans les codages α , β , γ . On a : $\delta_\Delta(P,P') = \delta_1(\vec{\alpha},\vec{\alpha}') = \frac{1}{2}\delta_1(\vec{\gamma},\vec{\gamma}') = \frac{1}{2}\delta_1(\vec{\beta},\vec{\beta}') = \delta_H(\vec{\alpha},\vec{\alpha}') = \delta_H(\vec{\gamma},\vec{\gamma}') = \delta_E^2(\vec{\alpha},\vec{\alpha}') = \frac{1}{4}\delta_E^2(\vec{\gamma},\vec{\gamma}')$.

Compte tenu de cette proposition, il n'y a donc que trois dissimilarités différentes associées aux trois codages des préordres totaux dans $\mathbb{R}^{n^2}$: δ_Δ, δ_H $(\vec{\beta},\vec{\beta}')$ et δ_E^2 $(\vec{\beta},\vec{\beta}')$. Les deux premières sont des distances; la première , i.e. la distance de la différence symétrique est bien connue; quant à la deuxième, on peut remarquer que divisée par deux, elle dénombre les paires $\{x,y\}$ pour lesquelles P et P' ont des restrictions différentes : $P_{\{x,y\}} \neq P'_{\{x,y\}}$. Puisque ces deux dissimilarités sont des distances, il en est de même de leur moyenne arithmétique dont on verra qu'il est intéressant de la considérer. Par contre δ_E^2 $(\vec{\beta},\vec{\beta}')$ n'est pas une distance (elle ne vérifie pas l'inégalité triangulaire). Les quatre dissimilarités précédentes permettent de définir quatre coefficients d'accord qui tous généralisent le tau de Kendall entre ordres totaux; pour cette raison, nous les notons : τ_1 , τ_2 , τ_3 et τ_4 . Ils sont étudiés dans le paragraphe 3.3. Dans ce même paragraphe, on considérera trois autres généralisa-

tions du tau , obtenues, l'une (τ_5) , à partir d'une dissimilarité qui n'est ni une distance, ni un carré de distance, les deux autres (τ_a et τ_F) à partir d'une pseudo-dissimilarité.

Au paragraphe 3.4, on étudiera une généralisation ρ_1 du rho de Spearman entre deux ordres totaux. Elle sera obtenue à partir du codage rang moyen des préordres totaux dans $\mathbb{R}^n$ (cf. 2.2 - *Codage des préordres totaux dans* $\mathbb{R}^n$), en utilisant le carré de la distance euclidienne $\delta_E(\vec{r},\vec{r}')$. Enfin, en conclusion on évoquera d'autres possibilités non étudiées ici.

3.3 - Coefficients d'accord généralisant le tau de Kendall

Nous allons définir 7 tels coefficients; afin de faciliter leur comparaison ultérieure (cf. 3.3 - *Synthèse*) nous les définissons à partir de dissimilarités qui ont toutes le même intervalle de variation : $[0,n(n-1)]$. Ceci nous amène éventuellement à faire une transformation linéaire des dissimilarités définies au paragraphe précédent, transformation qui évidemment, ne modifie pas le coefficient d'accord associé (cf. la formule (1) de 3.1). Nous commençons par présenter séparément chaque coefficient, puis en 3.3 - *Synthèse*, nous en donnons une présentation synthétique ainsi que les références concernant les coefficients déjà connus.

Le coefficient τ_1

Ce coefficient est obtenu en prenant comme dissimilarité $\delta_1(P,P')$ la distance de Hamming $\delta_H(\vec{\beta},\vec{\beta}')$, où $\vec{\beta}$ et $\vec{\beta}'$ sont les vecteurs codant P et P' dans le codage β (2.2 - *Codage des préordres totaux dans* $\mathbb{R}^{n^2}$). Un calcul simple montre qu'on a alors :
$\delta_1(P,P') = 2(|O \cap O'^r| + |O^r \cap I'| + |I \cap O'^r|) = \delta_\Delta(P,P') + |O^r \cap I'| + |I \cap O'^r|$ et on trouve aisément que le maximum de cette dissimilarité est $n(n-1)$ obtenue si et seulement si l'intersection de P et P'^r est un ordre total. La similarité $\sigma_1(P,P')$ associée à δ est donc :

$$\sigma_1(P,P') = 2|O \cap O'| + |\overset{o}{I} \cap \overset{o}{I}'|$$

On obtient alors les trois expressions suivantes de τ_1 (correspondant aux formules (1), (2), (3) de 3.1) :

$$\tau_1(P,P') = 1 - \frac{4(|O \cap O'^r| + |O^r \cap I'| + |I \cap O'^r|}{n(n-1)}$$

$$= \frac{4|O \cap O'| + 2|\overset{o}{I} \cap \overset{o}{I'}|}{n(n-1)} - 1$$

$$= \frac{2(|O \cap O'| - |O \cap O'^r|) + |\overset{o}{I} \cap \overset{o}{I'}| - 2(|O^r \cap I'| + |I \cap O'^r|}{n(n-1)}$$

En utilisant les notations de 2.3 - *Partitions et paramètres fondamentaux associés à une paire de préordres totaux*, on peut écrire, de façon condensée :

$$\tau_1 = \frac{2(a-d) + 2(c^*-b)}{n(n-1)} .$$

En particulier, on obtient immédiatement des formules pour tous les types de paires particulières de préordres totaux caractérisées en 2.3 par la nullité de certains de ces paramètres.

Pour l'exemple de référence, on obtient $\delta_1(P,P') = 26$ et
$\tau_1(P,P') = -.73$.

Le coefficient τ_2

On prend comme dissimilarité entre P et P' la quantité suivante :

$$\delta_2(P,P') = 2|O \cap O'^r| + \frac{3(|O^r \cap I'| + |I \cap O'^r|)}{2} .$$

On remarquera que cette quantité est la moyenne arithmétique entre les deux distances $\delta_H(\vec{\beta},\vec{\beta}')$ (cf. 3.3 - *Le coefficient* τ_1) et $\delta_\Delta(P,P')$ (cf. Proposition 1, 3.2). C'est donc une distance, dont le maximum n(n-1) est obtenu si et seulement si on a deux ordres totaux réciproques. La similarité associée est donc :

$$\sigma_2(P,P') = 2|O \cap O'| + \frac{|O \cap I'| + |I \cap O'|}{2} + |\overset{o}{I} \cap \overset{o}{I'}|$$

et le coefficient d'accord :

$$\tau_2(P,P') = 1 - \frac{(4|O \cap O'^r| + 3|O \cap I'| + |I \cap O'^r|)}{n(n-1)}$$

$$= \frac{4|O \cap O'| + |O \cap I'| + |I \cap O'| + 2|\overset{o}{I} \cap \overset{o}{I'}|}{n(n-1)} - 1$$

$$= \frac{2(|O \cap O'| - |O \cap O'^{r}|) + |\overset{o}{I} \cap \overset{o}{I'}| - |O^{r} \cap I'| - |I \cap O'^{r}|}{n(n-1)}$$

ou, en notation abrégée :

$$\tau_2 = \frac{2(a-d) + 2c^* - b}{n(n-1)} \quad .$$

Dans l'exemple de référence, on a $\delta_2(P,P') = 24.5$ et
$\tau_2(P,P') = -.63$.

Le coefficient τ_3

On prend cette fois comme dissimilarité δ_3 la distance de la différence symétrique δ_Δ , dont il est facile de montrer que la valeur maximum $n(n-1)$ est obtenue si et seulement si on a deux ordres totaux réciproques. Donc :

$$\delta_3(P,P') = 2|O \cap O'^{r}| + |O^{r} \cap I'| + |I \cap O'^{r}|$$

$$\sigma_3(P,P') = 2|O \cap O'| + |I \cap I'| + |O \cap I'| + |I \cap O'|$$

et

$$\tau_3(P,P') = 1 - \frac{(4|O \cap O'^{r}| + 2|O^{r} \cap I'| + 2|I \cap O'^{r}|}{n(n-1)}$$

$$= \frac{4|O \cap O'| + 2|\overset{o}{I} \cap \overset{o}{I'}| + 2|O \cap I'| + 2|I \cap O'|}{n(n-1)} - 1$$

$$= \frac{2(|O \cap O'| - |O \cap O'^{r}|) + |\overset{o}{I} \cap \overset{o}{I'}|}{n(n-1)}$$

et en notation abrégée :

$$\tau_3 = \frac{2(a-d) + 2c^*}{n(n-1)}$$

Pour l'exemple de référence, on a $\delta_\Delta(P,P') = 23$ et
$\tau_3(P,P') = -.53$.

Le coefficient τ_4

On prend maintenant comme dissimilarité entre P et P' , $\delta_4(P,P') = \frac{1}{4}\,\delta^2_{E_1}(\vec{\beta},\vec{\beta}')$. Nous laissons au lecteur le soin de montrer que :

$$\delta_4(P,P') = \frac{1}{4}\,\delta^2_E(\vec{\beta},\vec{\beta}') = 2|O \cap O'^r| + \frac{1}{2}\,(|O^r \cap I'| + |I \cap O'^r|)$$

et que le maximum de cette quantité est $n(n-1)$ obtenu si et seulement si on a deux ordres totaux réciproques. On en déduit :

$$\sigma_4(P,P') = 2|O \cap O'| + |\overset{o}{I} \cap \overset{o}{I'}| + \frac{3}{2}\,(|O \cap I'| + |O' \cap I|) .$$

D'où les expressions suivantes de τ_4 :

$$\tau_4(P,P') = 1 - \frac{(4|O \cap O'^r| + |O^r \cap I'| + |I \cap O'^r|)}{n(n-1)}$$

$$= \frac{4|O \cap O'| + 2|\overset{o}{I} \cap \overset{o}{I'}| + 3(|O \cap I'| + |O' \cap I|)}{n(n-1)} - 1$$

$$= \frac{2|O \cap O'| - |O \cap O'^r|) + |\overset{o}{I} \cap \overset{o}{I'}| + (|O \cap I'| + |I \cap O'|)}{n(n-1)}$$

et en abrégé :

$$\tau_4 = \frac{2(a-d) + 2c^* + b}{n(n-1)}$$

Pour l'exemple de référence, $\frac{1}{2}\,\delta^2_E(\vec{\beta},\vec{\beta}') = 43$ et

$$\tau_4(P,P') = -.43 .$$

Le coefficient τ_5

Ce coefficient est associé à la dissimilarité entre P et P' définie comme le nombre de leurs désaccords stricts. On a donc :

$$\tau_5(P,P') = 2|O \cap O'^r| = |O \cap P'^c| + |P^c \cap O'| .$$

Le maximum de cette dissimilarité étant $n(n-1)$ (obtenu si et seulement si P et P' sont deux ordres totaux réciproques) on en déduit que la similarité correspondante est :

$$\sigma_5(P,P') = 2(|O \cap O'| + |O \cap I'| + |O' \cap I|) + |\overset{o}{I} \cap \overset{o}{I'}|$$

Le coefficient τ_5 s'écrit donc :

$$\tau_5(P,P') = 1 - \frac{4|O \cap O'^r|}{n(n-1)}$$

$$= \frac{4(|O \cap O'| + |O \cap I'| + |O' \cap I|) + 2|\overset{o}{I} \cap \overset{o}{I'}|}{n(n-1)} - 1$$

$$= \frac{2(|O \cap O'| - |O \cap O'^r|) + |\overset{o}{I} \cap \overset{o}{I'}| + 2(|O \cap I'| + |O' \cap I|)}{n(n-1)}$$

Soit :

$$\tau_5 = \frac{2(a-d) + 2(c^*+b)}{n(n-1)}$$

Pour l'exemple de référence, $2|O \cap O'^r| = 20$ et

$$\tau_5(P,P') = -\frac{1}{3} .$$

Il est clair que la dissimilarité δ_5 n'est pas une distance. En particulier, elle est nulle et $\tau_5(P,P')$ est donc égal à $+1$ si seulement si $P \cap P'$ est un préordre total. Par contre, comme pour les 3 coefficients précédents, on a $\tau_5(P,P') = -1$ si et seulement si P et P' sont deux ordres totaux réciproques.

Le coefficient τ_a

On utilise la mesure de désaccord suivante :

$$\delta_a(P,P) = 2|O \cap O'^r| + |O^r \cap I'| + |O'^r \cap I| + \frac{1}{2}|\overset{o}{I} \cap \overset{o}{I'}| .$$

On remarque que δ_a est seulement une pseudo-dissimilarité puisque $\delta_a(P,P) = \frac{|\overset{o}{I}|}{2}$ n'est nul que si P est un ordre total. On vérifie que le maximum de δ_a est $n(n-1)$ obtenu si et seulement si P et P' sont deux ordres totaux réciproques. La pseudo-similarité correspondante est donc :

$$\sigma_a(P,P') = 2|O \cap O'| + |O \cap I'| + |I \cap O'| + \frac{1}{2}|\overset{o}{I} \cap \overset{o}{I'}|$$

et les trois expressions de τ_a sont :

$$\tau_a(P,P') = 1 - \frac{4|O \cap O'^r| + 2|O^r \cap I'| + 2|O'^r \cap I| + |\overset{o}{I} \cap \overset{o}{I'}|}{n(n-1)}$$

$$= \frac{4|0 \cap 0'| + 2|0 \cap I'| + 2|I \cap 0| + |\overset{o}{I} \cap \overset{o}{I'}|}{n(n-1)} - 1$$

$$= \frac{2(|0 \cap 0'| - |0 \cap 0'^{r}|)}{n(n-1)}$$

Soit :

$$\tau_a = \frac{2(a-d)}{n(n-1)}$$

Pour l'exemple de référence, on obtient :

$$\tau_a(P,P') = -.53 \ .$$

Le coefficient τ_F

On utilise la mesure de désaccord suivante :

$$\delta_F(P,P') = |\overset{o}{P} \Delta 0'| = |0 \Delta \overset{o}{P'}|$$

$$= 2|0 \cap 0'^{r}| + |0^{r} \cap I'| + |0'^{r} \cap I| + |\overset{o}{I} \cap \overset{o}{I'}| \ .$$

La quantité δ_F est une pseudo-dissimilarité puisque $\delta_F(P,P) = |\overset{o}{I}|$ n'est nulle que si P est un ordre total. Son maximum étant $n(n-1)$ obtenu si et seulement si P et P' sont deux préordres totaux réciproques. La pseudo-similarité correspondante est :

$$\sigma_F(P,P') = 2|0 \cap 0'| + |0 \cap I'| + |I \cap 0'| \ .$$

On en déduit les trois expressions de τ_F :

$$\tau_F(P,P') = 1 - \frac{4|0 \cap 0'^{r}| + 2|0^{r} \cap I'| + 2|0'^{r} \cap I| + 2|\overset{o}{I} \cap \overset{o}{I'}|}{n(n-1)}$$

$$= \frac{4|0 \cap 0'| + 2|0 \cap I'| + 2|I \cap 0'|}{n(n-1)} - 1$$

$$= \frac{2(|0 \cap 0'| - |0 \cap 0'^{r}|) - |\overset{o}{I} \cap \overset{o}{I'}|}{n(n-1)}$$

Soit :

$$\tau_F = \frac{2(a-d) - 2c^*}{n(n-1)}$$

Dans l'exemple de référence, on obtient :

$$\tau_F(P,P') = -.53 \quad .$$

Synthèse

Le tableau ci-après présente, sous forme synthétique, les sept coefficients τ_i, ainsi que les similarités et dissimilarités correspondantes, notées respectivement σ_i et δ_i.

	$\lvert o \cap o' \rvert + \lvert o^r \cap o'^r \rvert$	$\lvert o' \cap x \rvert + \lvert o \cap x' \rvert$	$\frac{\lvert \overset{o}{x} \cap \overset{o}{x}' \rvert}{2}$		$\frac{\lvert \overset{o}{x} \cap \overset{o}{x}' \rvert}{2}$	$\lvert o'^r \cap x \rvert + \lvert o^r \cap x' \rvert$	$\lvert o^r \cap o' \rvert + \lvert o \cap o'^r \rvert$
	2a	b	c*		c*	b	2d
σ_1	1	0	2	δ_1	0	2	1
σ_2	1	$\frac{1}{2}$	2	δ_2	0	$\frac{3}{2}$	1
σ_3	1	1	2	δ_3	0	1	1
σ_4	1	$\frac{3}{2}$	2	δ_4	0	$\frac{1}{2}$	1
σ_5	1	2	2	δ_5	0	0	1
σ_a	1	1	1	δ_a	1	1	1
σ_F	1	1	0	δ_F	2	1	1
τ_1	+1	0	+1		+1	-2	-1
τ_2	+1	0	+1		+1	-1	-1
τ_3	+1	0	+1		+1	0	-1
τ_4	+1	+1	+1		+1	0	-1
τ_5	+1	+2	+1		+1	0	-1
τ_a	+1	0	0		0	0	-1
τ_F	+1	0	-1		-1	0	-1

Pour obtenir l'expression d'une dissimilarité, on fait le "produit scalaire" de la première demi-ligne du tableau avec la demi-ligne correspondante. Par exemple, $\sigma_1 = 2a \times 1 + b \times 0 + c^* \times 2 = 2(a+c^*)$. On obtient de même l'expression d'une dissimilarité. Pour obtenir l'expression de τ_i on fait le produit scalaire de la première ligne du tableau par la ligne correspondante à τ_i et on divise par $n(n-1)$. Par exemple,

$$n(n-1)\tau_1 = 2a + 0 + c^* + c^* - 2b - 2d = 2(a - d + c^* - b) .$$

Commençons par donner quelques repères historiques sur les sept coefficients, en notant qu'ils peuvent apparaître sous les deux autres formes équivalentes des dissimilarités et similarités associées.

Le coefficient τ_3 est en un sens le plus ancien puisque sous sa forme dissimilarité ce n'est autre que la distance de la différence symétrique. Sous cette forme il a été proposé par de la Vega et Lerman pour comparer des préordonnances, à la suite de travaux de Regnier et de Benzécri (cf. Benzécri, 1973, pp. 239-242). On notera que compte tenu de l'égalité $\delta(O,O') = \delta(P,P')$ de la Proposition 1 et de celles de la Proposition 2 (3.2), ce coefficient peut être défini à partir de nombreux codages et distances différentes. D'autre part on sait qu'un préordre total P (comme toute relation binaire) définit une variable dichotomique pour chaque couple d'éléments de X : (x,y) appartient ou n'appartient pas à P . Le coefficient τ_3 , sous sa forme similarité, n'est alors autre que le classique coefficient de Sokal et Michener (1958) pour de telles variables (cf., par exemple, Chandon et Pinson, 1980).

Dans le cadre de ce qu'il appelle les "indices of configural similarity" entre différentes structures, Lingoes (1967, 1973) a proposé les coefficients τ_1 et τ_5 qu'il appelle respectivement m et m^* (cf. Lingoes et autres, 1979). La dissimilarité associée à τ_5 , a été aussi utilisée pour la comparaison de préordonnances en taxonomie (cf. notamment Lerman, 1970, où $\delta_5(P,P')$ est appelée le nombre d'inversions entre P et P'). Nous ne connaissons pas de référence où apparaissent les coefficients τ_2 , et τ_4 , deux coefficients qui sont en un sens symétriques (cf. ci-dessous).

Le coefficient τ_a est classique puisque c'est l'un des deux coefficients proposés par Kendall pour généraliser τ aux préordres totaux (cf. Kendall, 1945, 1970).

Le coefficient τ_F apparaît (sous sa forme dissimilarité) dans un texte de Falguerolles (1976), à propos de comparaison de préordonnances (il l'attribue d'ailleurs faussement à de la Véga).

On notera les relations d'ordres suivantes (évidentes à partir des formules) :

$$\tau_1 \leq \tau_2 \leq \tau_3 \leq \tau_4 \leq \tau_5 \qquad \tau_F \leq \tau_a \leq \tau_3 \ .$$

On a d'autre part les relations arithmétiques suivantes :

$$\tau_3 = \frac{\tau_1 + \tau_5}{2} \ ; \quad \tau_2 = \frac{\tau_1 + \tau_3}{2} = \frac{3\tau_1 + \tau_5}{4}$$

$$\tau_4 = \frac{\tau_3 + \tau_5}{2} = \frac{\tau_1 + 3\tau_5}{4} \ ; \quad \tau_a = \frac{\tau_3 + \tau_F}{2} \ .$$

En particulier, la dissimilarité proposée par Falguerolles (1976) comme moyenne de $\delta_\Delta(P,P')$ et $\delta_\Delta(O,P')$ correspond au coefficient τ_a . D'autre part τ_1 et τ_5 apparaissent comme deux extrêmes qui donnent par moyennage les trois autres coefficients τ_2, , τ_3 et τ_4 .

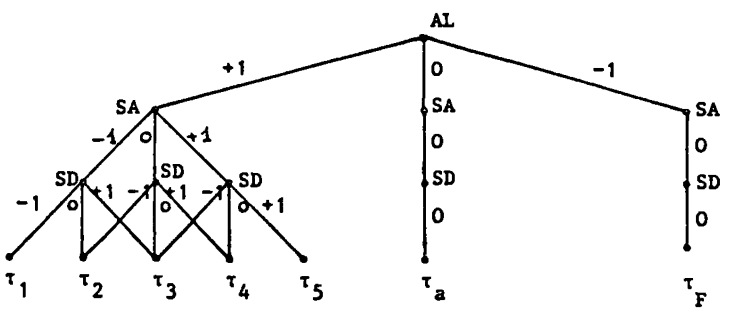
Si on considère notre exemple de référence, constitué par les deux préordres totaux $P = a < bc < df < e$ et $P' = f < de < b < a < c$, on constate que la valeur des sept coefficients τ_i varie de $-.73$ à $-.33$, ce qui est relativement important. Ces différences proviennent évidemment des différences de pondérations utilisées pour évaluer la contribution - positive ou négative - d'un couple (x,y) à la valeur du coefficient d'accord τ_i entre P et P' ; pour les sept coefficients ces poids prennent les valeurs 0 , ± 1 et ± 2 . Les choix faits entre ces poids sont explicités par les constations suivantes (cf. 3.1 et le tableau ci-dessus) :

1°) Les sept coefficients τ_i comptent tous $+1$ un accord strict ou un coaccord, et -1 un désaccord strict.

2°) Les cinq coefficients τ_1 à τ_5 comptent $+1$ accord large, tandis que τ_a ne le compte pas (poids 0), et que τ_F le compte -1 .

3°) Les coefficients τ_1 à τ_5 sont différenciés par les poids donnés aux semi-accords et aux semi-désaccords. Le coefficient τ_1 (τ_2) compte -2 (-1) un semi-désaccord et ne compte pas un semi-accord; à l'inverse τ_5 (τ_4) compte +2 (+1) un semi-accord et ne compte pas un semi-désaccord; enfin τ_3 ne compte ni les semi-accords ni les semi-désaccords.

Les constatations ci-dessus traduisent exactement les définitions - et donc les formules - des différents coefficients. On peut toutefois donner une autre interprétation à ces formules, en jouant du fait que les nombres de semi-accords et de semi-désaccords sont égaux. On peut en effet considérer que les poids donnés à n'importe quel couple sont toujours ± 1 ou zéro et qu'ils sont attribués selon le schéma arborescent suivant :



Dans cet arbre les noeuds non terminaux représentent un couple d'accord large, de semi-accord ou de semi-désaccord, et les valeurs sur les arêtes les pondérations données à de tels couples; les couples d'accord strict, de coaccords ou de désaccords stricts ne sont pas représentés puisque leur pondération est toujours la même. Les noeuds terminaux correspondent chacun à un des sept coefficients introduits. On remarque ainsi que le même coefficient peut s'obtenir à partir de différentes pondérations, c'est-à-dire donner lieu à différentes interprétations. On notera aussi que les branches

centrales et droites de cet arbre auraient pu être développées comme la branche gauche, ce qui aurait permis de définir huit autres coefficients. Au total on aurait ainsi quinze coefficients, généralisant tous le tau de Kendall. Le choix d'un de ces coefficients va finalement se ramener à faire un choix selon les deux critères indépendants suivants :

A. Pondération des accords larges

Il y a trois possibilités suivant que l'on veut que ces accords contribuent positivement, négativement, ou pas du tout au coefficient "global" d'accord entre P et P' . Le choix de la première possibilité paraît préférable si par exemple P et P' sont deux préordres totaux exprimant les préférences de deux sujets vis-à-vis d'objets à comparer.

B. Pondération des semi-accords et des semi-désaccords

On a ici cinq possibilités entre lesquelles le choix se fera suivant le contexte dans lequel se pose le problème de comparaison des préordres totaux considérés. Par exemple, dans un contexte de comparaison de préférences de deux sujets, il peut paraître naturel de ne pas compter les semi-accords et les semi-désaccords (puisqu'ils se correspondent); si d'autre part on a compté positivement les accords larges, ces choix conduisent au coefficient τ_3 (associé à la distance de la différence symétrique). Par contre, si on est dans un contexte de discrimination où il existe un préordre objectif qu'un sujet doit découvrir (par exemple au moyen de comparaisons par paires), on pourra estimer qu'un semi-accord doit être compté positivement; si on veut toutefois lui donner une importance moins grande qu'un accord strict, on pourra utiliser le coefficient τ_4 .

Remarque :

1. Si p est l'un quelconque des quatre paramètres fondamentaux a , b , c , d , associé à une paire de préordre totaux, on vérifie aisément que $p(P,P') = p(P^r,P'^r)$. On en déduit que tous les coefficients τ_i définis ci-dessus vérifient la condition $\tau_i(P,P') = \tau_i(P^r,P'^r)$.
2. Le coefficient de similarité de Russel et Rao pour des variables dichotomiques (cf. Chandon et Pinson, 1980) pourrait être utilisé pour comparer deux préordres totaux; c'est d'ailleurs celui proposé initialement par

Benzécri pour comparer deux préordonnances. Le coefficient d'accord (normalisé associé est $\frac{2(c^*-d)}{n(n-1)}$. On voit qu'il ne tient pas compte des accords stricts ! (en particulier, il ne généralise pas le tau de Kendall entre ordres totaux, puisqu'il ne prend dans ce cas que des valeurs négatives ou nulles).

3. Dans le tableau des coefficients τ_i , on voit que ceux-ci s'obtiennent comme combinaison linéaire des nombres a , b , c^* , d , les poids utilisés appartenant toujours à l'ensemble $\{-2,-1,0,+1,+2\}$. Il est clair qu'on pourrait utiliser d'autres pondérations, adaptées au contexte précis d'un problème de comparaison. On remarque aussi que dans le cas des semi-désaccords, on a donné toujours le même poids aux couples de $O'^r \cap I$ et à ceux de $O^r \cap I'$ (et de même pour les semi-accords correspondants). Dans le cas évoqué ci-dessus où P est un préordre objectif, on peut toutefois estimer qu'il est plus "grave" de discriminer deux objets identiques objectivement que de confondre deux objets distincts objectivement; dans ce cas on serait amené à donner des poids différents aux nombres b_1 et b_2 de semi-désaccords de ces deux types (les coefficients d_a et d_b définis au chapitre 4 sont des exemples de tels coefficients).

3.4 - Le coefficient ρ_1 (généralisant le rho de Spearman)

On utilise comme dissimilarité entre préordres totaux la quantité :

$$\delta(P,P') = \delta_E^2(\vec{r},\vec{r}')$$

où $\vec{r}$ et $\vec{r}'$ sont les vecteurs "rang moyen" correspondants à P et P' . On obtient ainsi un coefficient d'accord, noté ρ_1 , et qui s'écrit

$$\rho_1(P,P') = 1 - \frac{2\,\delta_E^2(\vec{r},\vec{r}')}{\max \delta_R^2(\vec{r},\vec{r}')}$$

Proposition

Max $\delta_E^2(\vec{r},\vec{r}') = \frac{n^3 - n}{3}$, ce maximum étant atteint si et seulement si P et P' sont deux ordres totaux réciproques.

Pour démontrer cette proposition nous considérons deux préordres totaux P et P', de vecteurs rangs $\vec{r}$ et $\vec{r}'$ et nous supposons $\delta^2_E(\vec{r},\vec{r}')$ maximum. On peut toujours supposer (en numérotant de façon convenable les éléments de X) $r_1 \leq r_2 \leq \dots \leq r_i \leq \dots \leq r_n$. Supposons maintenant qu'il existe $k < \ell$ avec $r'_k < r'_\ell$. Notons $\vec{s}$ le vecteur défini par :

pour tout $i \neq k,\ell$ $\quad s_i = r'_i \ ; \ s_k = r'_\ell \ ; \ s_\ell = r'_k$

$\vec{s}$ est le vecteur rang du préordre total obtenu à partir de P' en échangeant" k et ℓ. On a

$$\delta^2_E(\vec{r},\vec{s}) - \delta^2_E(\vec{r},\vec{r}') = (r_k-r'_\ell)^2 + (r_\ell-r'_k)^2 - (r_k-r'_k)^2 - (r_\ell-r'_\ell)^2 =$$

$$= 2(r_\ell-r_k)(r'_\ell-r'_k) \geq 0 .$$

On peut donc supposer le préordre P' tel que $r'_1 \geq r'_2 \geq \dots r'_i \geq \dots r'_n$.

Supposons maintenant que le vecteur $\vec{r}$ ait des coordonnées égales; pour simplifier les notations, on supposera que ses m premières coordonnées sont égales, donc que, pour $m > 1$, on ait :

$$r_1 = r_2 = \dots r_m = \frac{m+1}{2} < r_{m+1} \leq \dots \leq r_n .$$

Notons $\vec{s}$ le vecteur rang obtenu en réordonnant totalement ces m premiers éléments, de façon à avoir :

$$s_1 = 1, \dots s_m = m \quad ; \quad s_i = r_i \text{ pour tout } i \geq m+1 .$$

Un calcul simple donne

$$\delta^2_E(\vec{s},\vec{r}') - \delta^2_E(\vec{r},\vec{r}') = \frac{m^3-m}{12} + \sum_{i=1}^{m} r'_i \left(\frac{m+1}{2} - i\right) .$$

Dans cette somme de deux termes, le premier est toujours strictement positif. Quant au second, on vérifie qu'il s'écrit aussi, en posant $m = 2k$ ou $2k+1$

$$\sum_{j=0}^{k-1} (m-2j-1)(r'_{j+1} - r'_{m-j}) .$$

Compte tenu des inégalités sur les r'_i, ce deuxième terme est donc toujours positif ou nul. On obtient donc $\delta^2_E(\vec{s},\vec{r}') > \delta^2_E(\vec{r},\vec{r}')$ ce qui est im-

possible; on a donc $r_1 < r_2 \ldots < r_i \ldots < r_n$. Un raisonnement analogue montre qu'on a nécessairement $r'_1 > r'_2 > \ldots r'_i > \ldots r'_n$. P et P' sont donc deux ordres totaux réciproques et un calcul évident permet de conclure.

De la Proposition précédente on déduit l'expression suivante de ρ_1 :

$$\rho_1(P,P') = 1 - 6 \frac{\delta_E^2(\vec{r},\vec{r}')}{n^3 - n}$$

Pour l'exemple de référence on a $\delta_E^2(\vec{r},\vec{r}') = 59$ et $\rho_1(P,P') = -.69$.

Remarque :

Dans le cas d'ordres totaux, Spearman (1904) avait proposé une méthode rapide de calculer un coefficient de corrélation entre rangs, basée sur la distance de la norme L_1 :

$\delta_1(\vec{r},\vec{r}') = \sum_{i=1}^{n} |r_i - r'_i|$, normalisée par l'espérance mathématique de cette distance. Cette "footrule" est aussi discutée dans Kendall. Si on veut normaliser par rapport au maximum de cette distance, il faut le calculer, ce qui est fait dans Diaconis et Graham (1977), qui déterminent aussi les couples d'ordres totaux correspondants. Ces derniers résultats sont généralisés aux préordres totaux par Giakoumakis (1985), qui utilise la distance de la norme L_1 entre les vecteurs rangs moyens. Il montre que le maximum de cette distance est $[\frac{n^2}{2}]$ (comme dans le cas des ordres totaux), et que ce maximum est réalisé pour les couples (P,P') de préordres totaux obtenus en bipartitionnant de manière "équilibrée" l'ensemble X ($A+\bar{A} = X, |A|-|\bar{A}| \leq 1$) puis en définissant P (respectivement P'), comme un préordre arbitraire sur A (respectivement $\bar{A}$) suivi d'un préordre total arbitraire sur $\bar{A}$ (respectivement A) . Nous n'avons pas retenu ici le coefficient associé à cette distance puisqu'il ne généralise pas rho .

3.5 - Conclusion

L'approche utilisée dans cette troisième partie pour définir un coefficient d'accord entre préordres totaux, i.e. la normalisation d'une dissi-

milarité (ou d'une similarité), nous a permis de retrouver certains coefficients plus ou moins classiques (τ_a de Kendall, m et m* de Lingoes, τ_3 de Sokal et Michener, de la Vega et Lerman, etc...), d'en introduire d'autres tout aussi admissibles que les précédents, et surtout de les situer tous dans un même cadre, ce qui rend plus aisé (mais non nécessairement plus facile) leur comparaison. En particulier, on voit que cette comparaison se ramène à celle des dissimilarités qui permettent de les définir.

Pour terminer il est donc opportun de souligner que d'autres dissimilarités que celles utilisées ici auraient pu être considérées. Citons par exemple les métriques "laticielles" associées à la structure de demi-treillis des préordres totaux (l'étude d'un coefficient d'accord associée à une telle métrique est faite dans Giakoumakis 1985). Citons également les distances de la différence symétrique "pondérée" ou plus généralement associées à des valuations "supérieures" ou "inférieures" (telle que l'entropie) définies sur l'ensemble des préordres totaux; on en trouvera une étude détaillée dans Barthélémy (1979). Pour obtenir un coefficient d'accord normalisé à partir de telles distances, il faudra évidemment calculer leur maximum ce qui n'est pas nécessairement simple.

4 - COEFFICIENTS D'ACCORD DEFINIS PAR LA NORMALISATION D'UN PRODUIT SCALAIRE

4.1 - Introduction

Les coefficients d'accord présentés dans cette section sont obtenus en normalisant le produit scalaire de deux vecteurs codant ces préordres totaux. Un premier paramètre définissant un tel coefficient est donc le codage utilisé. Celui-ci peut être l'un des codages définis en 2.1 ; mais on utilise aussi un codage β_p généralisant le codage β et que nous définirons ci-dessous. Un deuxième paramètre est le type de normalisation choisi, étant entendu que celle-ci fera varier le coefficient d'accord entre -1 et +1 . Une possibilité consiste à prendre le coefficient de corrélation linéaire entre les deux vecteurs, soit si on les note $\vec{x}$ et $\vec{y}$

$$\text{corr}\ (\vec{x},\vec{y}) = \frac{(\vec{x}-\vec{m}_x).(\vec{y}-\vec{m}_y)}{\|\vec{x}-\vec{m}_x\|\ \|\vec{y}-\vec{m}_y\|} .$$

Dans ce cas, nos codages étant toujours injectifs, le coefficient d'accord égale $+1$ si et seulement si $P = P' \neq X^2$. (Dans le cas $P = P' = X^2$, on a des vecteurs constants et donc indétermination).

D'autres possibilités de normalisation existent qui ont été essentiellement utilisées avec le codage β_p (ou son cas particulier β). Nous allons donc maintenant présenter ce codage et le principe de ces normalisations.

On se donne une pondération des couples de X^2 ; c'est-à-dire pour chaque couple (x,y) un nombre $p(x,y) \geq 0$. On définit le codage β_p du préordre total $P = O + I$ par

$$\begin{aligned}
\beta_p(x,y) &= +\sqrt{p(x,y)} && \text{si et seulement si} \quad (x,y) \in O \\
\beta_p(x,y) &= -\sqrt{p(x,y)} && \text{si et seulement si} \quad (x,y) \in O^r \\
\beta_p(x,y) &= 0 && \text{si et seulement si} \quad (x,y) \in I
\end{aligned}$$

On voit qu'on retrouve le codage β en prenant $p(x,y) = 1$ pour tout x,y , et que plus généralement si $p(x,y)$ est constant on a une transformation linéaire de ce codage.

Notons $\vec{\beta}_p$ le vecteur associé au préordre P, $\vec{\beta}'_p$ le vecteur associé au préordre P' . On pose

$$\Gamma_p(P,P') = \vec{\beta}_p.\vec{\beta}'_p = \sum_{x,y} p(x,y)\ \beta(x,y)\ \ \beta'(x,y) .$$

Cette quantité est une mesure "brute" de la concordance entre P et P' . Si on pose $W = \sum_{x \neq y} p(x,y)$, il est clair (puisque $\beta(x,y)\beta'(x,y) \in \{+1,0,-1\}$) que Γ_p varie de $-W$ à $+W$, valeurs obtenues si (mais, non généralement, seulement si), P et P' sont deux ordres totaux, soit réciproques, soit identiques. En particulier, dans le cas d'une pondération constante k , on obtient

$$\Gamma_p(P,P') = 2k\ (|O \cap O'| - |O \cap O'^r|)$$

A un coefficient près, cette quantité est donc le nombre d'accords "stricts" moins le nombre de désaccords "stricts" entre P et P' (où "strict" signifie qu'on ne considère que des couples n'appartenant pas à $I \cup I'$).

Les coefficients d'accord associés au produit scalaire Γ_p sont définis en normalisant cette quantité pour la faire varier de -1 à $+1$, ces deux valeurs étant obtenues si l'on a respectivement deux ordres totaux réciproques ou deux ordres totaux identiques. Il suffit pour cela de la diviser par un dénominateur positif $\Delta(P,P')$ tel que :

$$|\Gamma_p(P,P')| \leq \Delta(P,P') \leq W .$$

En fait, plusieurs solutions sont possibles, qui donnent chacune des caractéristiques différentes des couples des préordres totaux pour lesquels le coefficient correspondant prend ses valeurs extrêmes ± 1 .

Nous étudions les coefficients d'accord obtenus par différents choix de codages et de normalisations en distinguant ceux généralisant tau (4.2) de ceux généralisant rho (4.3) . Pour les premiers, nous en considérons d'abord six, définis par diverses normalisations de la quantité $\Gamma(P,P')$ égale au produit scalaire $\vec{\beta}\vec{\beta}'$ (et donc à la différence entre les nombres d'accords stricts et de désaccords stricts). Puis nous présentons les coefficients basés sur la normalisation de Γ_p , parfois appelés coefficients de monotonicité de Guttman. Enfin nous terminons par l'utilisation du coefficient de corrélation qui conduit soit au coefficient τ_b de Kendall (avec le codage β), soit à un nouveau coefficient τ_γ . Les coefficients généralisant rho étudiés en 4.3 sont les deux coefficients ρ_a et ρ_b de Kendall; ils sont obtenus en normalisant Γ_p , ou pour ρ_b en prenant le coefficient de corrélation linéaire associé au codage rang moyen (ou score).

4.2 - Coefficients généralisant le tau de Kendall

Les coefficients τ_a , e , τ_b , d_a , d_b , γ

La mesure "brute" d'accord entre P et P' est ici

$$\Gamma(P,P') = 2(|O \cap O'| - |O \cap O'^r|)$$

Pour la normaliser on va utiliser les six dénominateurs suivants :

$$\Delta_1 = n(n-1)$$

$$\Delta_2 = \sum_{x,y} (|\beta(x,y)| + |\beta'(x,y)| - |\beta(x,y)\beta'(x,y)|)$$

$$= 2[|O \cap O'| + |O \cap O'^r| + |I \cap O'| + |I' \cap O|] = n^2 - |I \cap I'|$$

$$\Delta_3 = (\sum_{x,y} |\beta(x,y)|)^{\frac{1}{2}} (\sum_{x,y} |\beta'(x,y)|)^{\frac{1}{2}} = \|\vec{\beta}\| \, \|\vec{\beta}'\| = 2(|O| \, |O'|)^{\frac{1}{2}}$$

$$\Delta_4 = \sum_{x,y} |\beta(x,y)| = \|\vec{\beta}\|^2 = 2|O| = n^2 - |I|$$

$$\Delta'_4 = \sum_{x,y} |\beta'(x,y)| = \|\vec{\beta}'\|^2 = 2|O'| = n^2 - |I'|$$

$$\Delta_5 = \sum_{x,y} |\beta(x,y)\beta'(x,y)| = 2(|O \cap O'| + |O \cap O'^r|) = n^2 - |I \cap I'|$$

On obtient ainsi les coefficients suivants (les noms entre parenthèses sont ceux des auteurs à qui ces coefficients sont attribués).

$$\tau_a = \frac{\Gamma}{\Delta_1} = \frac{2(|O \cap O'| - |O \cap O'^r|)}{n(n-1)} \qquad \text{(Kendall)}$$

$$e = \frac{\Gamma}{\Delta_2} = \frac{|O \cap O'| - |O \cap O'^r|}{|O \cap O'| + |O \cap O'^r| + |I \cap O'| + |I' \cap O|} \qquad \text{(Wilson)}$$

$$\tau_b = \frac{\Gamma}{\Delta_3} = \frac{|O \cap O'| - |O \cap O'^r|}{(|O \cap O'|)^{\frac{1}{2}}} \qquad \text{(Kendall)}$$

$$d_a = \frac{\Gamma}{\Delta_4} = \frac{|O \cap O'| - |O \cap O'^r|}{|O|} \; ; \; d_b = \frac{\Gamma}{\Delta'_4} = \frac{|O \cap O'| - |O \cap O'^r|}{|O'|} \qquad \text{(Sommers)}$$

$$\gamma = \frac{\Gamma}{\Delta_5} = \frac{|O \cap O'| - |O \cap O'^r|}{|O \cap O'| + |O \cap O'^r|} \qquad \text{(Goodman et Kruskal)}$$

Remarquons d'abord que le coefficient τ_a de Kendall a déjà été considéré en 3.3 comme normalisation d'une pseudo-dissimilarité; quant au coefficient τ_b de Kendall nous le réobtiendrons ci-dessous comme coefficient de corrélation. Notons aussi que les coefficients d_a et d_b de Sommers ne sont pas des coefficients d'accords au sens où nous les avons définis puisqu'ils ne sont pas symétriques : $d_a(P'P) = d_b(P,P') \neq d_a(P,P')$ (sauf si $|O| = |O'|$

et en particulier si P et P' sont des ordres totaux). En fait, d_a et d_b proviennent d'une "désymétrisation" de τ_b, puisque on a

$\tau_b = (d_a d_b)^{\frac{1}{2}}$.

Il est facile de vérifier que si P et P' sont des ordres totaux tous ces coefficients sont égaux à $2\frac{(|O \cap O'| - |O \cap O'^r|)}{n(n-1)}$, c'est-à-dire au tau de Kendall.

Il est d'autre part clair d'après les expressions des Δ_i qu'on a toujours

$$\Delta_1 \geq \Delta_2 \geq \Delta_3 \ , \ \Delta_4 \ , \ \Delta_4' \geq \Delta_5 \ .$$

Plus précisément Δ_i (pour $i \neq 3$) dénombre un ensemble de couples d'éléments de X obtenu en supprimant certains couples de $I \cup I'$. Pour Δ_i seuls les n couples (x,x) sont supprimés; pour Δ_2 on supprime les couples de $I \cap I'$: pour Δ_4 (Δ_4') ceux de $I(I')$; enfin pour Δ_5 on supprime tous les couples de $I \cup I'$.

Il résulte des inégalités ci-dessus, qu'on a toujours

$$|\tau_a| \leq |e| \leq |\tau_b| \ , \ |d_a| \ , \ |d_b| \leq |\gamma| \ .$$

Dans l'exemple de référence $P = 1 < 23 < 45 < 6$, $P' = 5 < 46 < 2 < 1 < 3$, on obtient $\gamma = -.67 < d_b = -.61 < \tau_b = -.59 < d_a = -.57 < e = \tau_a = -.53$.

Les inégalités ci-dessus permettent d'appréhender certains effets induits par le choix des normalisation sur la valeur des coefficients d'accord correspondants. Par exemple, pour P et P' fixé avec $|O \cap O'| - |O \cap O'^r| < 0$, on diminue la valeur du coefficient en choisissant des dénominateurs variant de τ_5 à τ_1 , c'est-à-dire prenant en compte de plus en plus de couples de $I \cup I'$! Malheureusement si pour P fixé on fait varier P' il est beaucoup moins simple de suivre et comparer l'évolution de ces coefficients puisque (sauf τ_a qui a un dénominateur fixe) ils sont le quotient de quantités variables mais reliées entre elles. Par exemple, la même modification de P' pourra laisser constante, augmenter ou diminuer les valeurs de ces coefficients.

Nous nous contenterons donc ici pour comparer ces coefficients d'en donner quelques propriétés simples portant sur leurs valeurs extrêmes ou les valeurs de couples (P,P') particuliers.

Le tableau ci-dessous caractérise les couples (P,P') de préordres totaux pour lesquels la valeur du coefficient indiqué est $+1$ ou -1 .

	τ_a	e, τ_b	d_a	d_b	γ
$+1$	$P = P' =$ ordre total	$P = P' \neq X^2$	$P' \subseteq P \neq X^2$	$P \subseteq P' \neq X^2$	$P \cap P' =$ préordre total $\neq X^2$
-1	$P = P'^r$ ordre total	$P = P'^r \neq X^2$	$P' \subseteq P^r \neq X^2$	$P \subseteq P'^r \neq X^2$	$P \cap P'^r =$ préordre total $\neq X^2$

On constate ainsi que si $P = P'$, ces coefficients (sauf $\tau_a = \frac{2|O \cap O'|}{n(n-1)}$ prennent la valeur $+1$; mais inversement la valeur $+1$ du coefficient peut signifier une simple inclusion (dans le cas de d_a et d_b), ou même seulement l'existence d'un ordre total inclus dans P et P' (cas de γ). On a des constations duales pour le cas de $P = P'^r$.

Pour un couple (P,P') avec $P \subseteq P'$, on a

$$\tau_a = \frac{2|O'|}{n(n-1)} \leq e = d_a = \frac{|O'|}{|O|} \leq \tau_b = \left(\frac{|O'|}{|O|}\right)^{\frac{1}{2}} < \gamma = d_b = +1$$

Notons enfin que ces six coefficients sont nuls si et seulement si le nombre d'accords stricts égale le nombre de désaccords stricts et que leur dénominateur est non nul (i.e. P ou P' différents de X^2 pour e , et P et P' différents de X^2 pour τ_b , d_a , d_b et γ).

Terminons ce paragraphe par deux remarques.

1. On peut donner (cf. par exemple Kruskal, 1958, Hubert et autres, 1985) une interprétation probabiliste des coefficients τ_a , e , d_a , d_b et γ . Ces coefficients s'écrivent en effet sous la forme $\frac{2|O \cap O'|}{\Delta} - \frac{2|O \cap O'^r|}{\Delta}$. Pour τ_a , $\Delta = n(n-1)$ et en supposant qu'on

tire au hasard une paire d'objets $\{x,y\}$ parmi les $\binom{n}{2}$ paires possibles, τ_a s'interprète comme la différence entre la probabilité d'avoir un accord strict sur $\{x,y\}$ moins la probabilité d'avoir un désaccord strict. Pour les autres coefficients, on a la même interprétation mais en supposant que le tirage au hasard se fait parmi un ensemble restreint de paires : les paires qui ne sont pas "ex-aequo" dans $P \cap P'$ pour e, celles qui ne sont pas "ex-aequo" dans $P(P')$ pour d_a (d_b), enfin celles qui ne sont ex-aequo ni dans P ni dans P', pour γ.

2. Le fait que la quantité Γ est le produit scalaire des vecteurs $\vec{\beta},\vec{\beta}$, et que les dénominateurs Δ_i s'expriment en fonction des coordonnées de ces vecteurs permet de généraliser les coefficients d'accords précédents à des structures plus générales que celles de préordres totaux. En particulier, Hubert et autres (1985) étudient le cas où les coordonnées du vecteur $\vec{\beta}$, prennent toujours les valeurs $0,\pm1$, mais où celui-ci code des k-uples d'éléments de X (au lieu de coder des couples).

Les coefficients de monotonicité de Guttman

Ces coefficients utilisent comme mesure brute d'accord la quantité $\Gamma_p(P,P') = \Sigma p(x,y)\ \beta(x,y)\ \beta'(x,y)$. Pour la normaliser on utilise les quatre dénominateurs suivants :

$$\Delta^{(1)} = \sum_{x,y} p(x,y)|\beta(x,y)\ \beta'(x,y)|$$

$$\Delta^{(2)} = \sum_{x,y} p(x,y)|\beta(x,y)| = \|\vec{\beta}_p\|^2$$

$$\Delta^{(3)} = \sum_{x,y} p(x,y)\ |\beta'(x,y) = \|\vec{\beta}'_p\|^2$$

$$\Delta^{(4)} = \sum_{x,y} p(x,y)\ (|\beta(x,y)|+|\beta'(x,y)|-|\beta(x,y).\beta'(x,y)|)$$

On pose ensuite pour $m = 1,2,3,4$

$$\mu^m(P,P') = \frac{\Gamma_p(P,P')}{\Delta^{(m)}}$$

Pour $m = 1$ à 4, $\mu^{(m)}$ est appelé respectivement le coefficient de mono-

tonicité faible, semi-fort, semi-faible, fort (de Guttman). On remarque que dans le cas particulier de la pondération constante, on retrouve respectivement les coefficients γ, d_a, d_b et e du paragraphe précédent. On retrouverait τ_a, en introduisant le dénominateur

$$\Delta^{(5)} = \sum_{x \neq y} p(x,y) .$$

Les coefficients de corrélation linéaire τ_b *et* τ_γ

Considérons d'abord le codage β des préordres totaux, et le coefficient de corrélation linéaire associé. Un calcul évident donne

$$\text{corr}\,(\vec{\beta},\vec{\beta}') = \frac{|O \cap O'| - |O \cap O'^{r}|}{(|O \cap O'|)^{\frac{1}{2}}}$$

On retrouve donc le coefficient τ_b déjà considéré ci-dessus et introduit par Kendall comme une normalisation possible de la différence entre accords stricts et désaccords stricts dans le cas des préordres totaux (cf. Kendal, 1945, 1970).

Utilisons maintenant le codage γ des préordres totaux et définissons un coefficient d'accord en posant

$$\tau_\gamma(P,P') = \text{corr}\,(\vec{\gamma},\vec{\gamma}')$$

Compte tenu des égalités $\vec{\gamma}\vec{\gamma}' = n(n-1) - 2d_\Delta(P,P')$, $\sum_{X^2} \gamma(x,y) = |I|^{o}$

$\sum_{X^2} \gamma^2(x,y) = n(n-1)$, on obtient finalement

$$\tau_\gamma(P,P') = \frac{n^3(n-1) - 2n^2\, d_\Delta(P,P') - |I|^{o}\,|I'|^{o}}{(n^3(n-1) - |I|^{2o})^{\frac{1}{2}}\,(n^3(n-1) - |I'|^{2o})^{\frac{1}{2}}}$$

Ce coefficient ayant été défini comme coefficient de corrélation linéaire est égal à $+1$ si et seulement si $P = P' \neq X^2$. De plus on montre facilement les résultats suivants :

- $\tau_\gamma(P,P') = -1$ si et seulement si P et P' sont deux ordres totaux réciproques.

$P = X^2$ et P' ordre total impliquent $\tau_\gamma = 0$.

• $P' = P^r$ implique $\tau_\gamma(P,P') = 1 - \dfrac{4n^2\ |\overset{o}{I}|}{n^3(n-1) - |\overset{o}{I}|^2}$

• P et P' ordres totaux impliquent $\tau_\gamma = \tau$.

Exemple

Dans l'exemple de référence, on a $d_\Delta(P,P') = 23$, $|\overset{o}{I}| = 4$, $|\overset{o}{I'}| = 2$, d'où $\tau_\gamma(P,P') = -.55$.

Remarque :

Puisque nous avons défini un troisième codage α des préordres totaux on aurait pu considérer le coefficient de corrélation linéaire correspondant $corr(\vec{\alpha},\vec{\alpha}')$. Ce coefficient est toutefois peu intéressant; en particulier si P et P' sont des ordres totaux, il est égal à $\dfrac{n\tau + 1}{n+1}$ et présente donc un biais par rapport au tau de Kendall. (Pour plus de détails voir Giakoumakis, 1985).

4.3 - Les coefficients généralisant le rho de Spearman : ρ_a et ρ_b

On définit ces coefficients en normalisant la quantité Γ_p , où la pondération p est donnée par :

$$p(x,y) = |r(x) - r(y)|.|r'(x) - r'(y)|$$

Ici $r(x)$ $(r'(x))$ est le rang moyen de x dans P (P') . Le codage associé est donc

$\beta_p(x,y) = (|r(x)-r(y)|.|r'(x)-r'(y)|)^{\frac{1}{2}}$ si et seulement si $(x,y) \in O$

$\beta_p(x,y) = 0$ si et seulement si $(x,y) \in I$

$\beta_p(x,y) = -(|r(x)-r(y)|.|r'(x)-r'(y)|)^{\frac{1}{2}}$ si et seulement si $(y,x) \in O$.

Puisque, par exemple, $r(x) < r(y)$ si et seulement si $(x,y) \in O$, et $r(x) = r(y)$ si et seulement $(x,y) \in I$, on en déduit que le produit scalaire des vecteurs $\vec{\beta}_p$ et $\vec{\beta}'_p$ est

$$\Gamma_p(P,P') = \vec{\beta}_p.\vec{\beta}'_p = \sum_{x,y} (r(x)-r(y))(r'(x)-r'(y)) .$$

Pour normaliser cette quantité, on utilise les deux dénominateurs suivants :

$$D_1 = \frac{n^3 - n}{6}$$

$$D_2 = \left(\sum_{x,y} [r(x)-r(y)]^2 \, \Sigma[r'(x)-r'(y)]^2 \right)^{\frac{1}{2}}$$

On obtient ainsi deux coefficients dus à Kendall (1945, 1970) :

$$\rho_a(P,P') = \frac{6 \sum_{x,y} [r(x)-r(y)][r'(x)-r'(y)]}{n^3 - n}$$

$$\rho_b(P,P') = \frac{\sum_{x,y} [r(x)-r(y)][r'(x)-r'(y)]}{\left(\sum_{x,y} [r(x)-r(y)]^2 \sum_{x,y} [r'(x)-r'(y)]^2 \right)^{\frac{1}{2}}}$$

L'expression ci-dessus de ρ_b montre que ce coefficient peut être défini comme un coefficient de corrélation linéaire. Il suffit pour cela de considérer le codage ε d'un préordre total dans $\mathbb{R}^{n^2}$ défini par $\varepsilon(x,y) = r(x) - r(y)$. Mais on peut aussi montrer que ρ_b est le coefficient de corrélation linéaire pour les vecteurs rang moyen :

$$\rho_b(P,P') = \text{corr}\,(\vec{r},\vec{r}')$$

Il suffit pour cela d'utiliser le lemme suivant :

Lemme

Soient $\vec{a} = (a_1,\ldots a_i,\ldots a_n)$, $\vec{a}' = (a'_1,\ldots a'_i,\ldots a'_n)$ deux vecteurs de $\mathbb{R}^n$ et $\vec{b} = (b_{11},\ldots,b_{ij},\ldots b_{nn})$, $\vec{b}' = (b'_{11},\ldots b'_{ij},\ldots b'_{nn})$ les deux vecteurs de $\mathbb{R}^{n^2}$ obtenus en posant $b_{ij} = a_i - a_j$, $b'_{ij} = a'_i - a'_j$. On a

$\text{cov}\,(\vec{b},\vec{b}') = 2n \,\text{cov}\,(\vec{a},\vec{a}')$; $\text{var}\,\vec{b} = 2n \,\text{var}\,\vec{a}$; $\text{corr}\,(\vec{b},\vec{b}') = \text{corr}\,(\vec{a},\vec{a}')$.

On a , en effet

$$\vec{b}.\vec{b} = \sum_{i,j} b_{ij} b'_{ij} = \sum_{i,j} (a_i - a_j).(a'_i - a'_j) = 2n \sum_i a_i a'_i - 2 \sum_{i,j} a_i a'_j =$$

$$= 2n \,(\vec{a}-\vec{m}_a).(\vec{a}'-\vec{m}'_a) \ .$$

On en déduit $\|\vec{b}\|^2 = 2n\|(\vec{a}-\vec{m}_a)\|^2$, et en remarquant que $\vec{b}$ et $\vec{b}'$ sont des vecteurs centrés, on obtient les égalités du lemme.

En utilisant ce même lemme, on montre aussi :

$$\rho_a(P,P') = 2\rho_b(P,P')\frac{(\text{var}\,\vec{r}\,.\,\text{var}\,\vec{r}')^{\frac{1}{2}}}{n^2-1} = \frac{12\,\text{cov}\,(\vec{r},\vec{r}')}{n^2-1}$$

Kendall donne des formules de calcul pour ρ_a et ρ_b que nous présentons ci-dessous. Posons

$P = C_1 < \ldots C_i \ldots < C_t$, $n_i = |C_i|$, pour $i = 1,\ldots t$;

$P' = C'_1 < \ldots C'_i \ldots < C'_u$, $n'_i = |C'_i|$, pour $i = 1,\ldots u$;

$$T = \sum_{i=1}^{t} \frac{(n_i^3 - n_i)}{2} \qquad U = \sum_{i=1}^{u} \frac{(n_i'^3 - n_i)}{2}$$

On a

$$\rho_a(P,P') = \frac{n^3-n-6\delta_E^2(\vec{r},\vec{r}')-(T+U)}{n^3-n}$$

$$\rho_b(P,P') = \frac{n^3-n-6\delta_E^2(\vec{r},\vec{r}')-(T+U)}{(n^3-n-2T)^{\frac{1}{2}}(n^3-n-2U)^{\frac{1}{2}}}$$

Dans le cas où P et P' sont des ordres totaux, on a $T = U = 0$ et ces deux coefficients redonnent donc le rho de Spearman entre ordres totaux.

Aux propriétés de ces coefficients qu'on trouve dans Kendall (1970), on peut ajouter les suivantes :

$\rho_a(P,P') = +1$ si et seulement si $P = P'$ est un ordre total

$\rho_a(P,P') = -1$ si et seulement si $P = P'^r$ est un ordre total

$\rho_b(P,P') = +1$ si et seulement si $P = P' \neq \chi^2$

$\rho_b(P,P') = -1$ si et seulement si $P = P'^r \neq \chi^2$.

Les résultats sur les bornes de ρ_a utilisent le lemme suivant (Giakoumakis, 1985).

Lemme

Soit $\vec{r}$ le vecteur rang moyen d'un préordre total. On a

$0 \leq \text{var } \vec{r} \leq \frac{n^2-1}{12}$, la borne supérieure étant atteinte si et seulement si P est un ordre total.

On déduit immédiatement de ce lemme et de la relation écrite plus haut entre ρ_a et ρ_b :

Pour tout P,P' $|\rho_a(P,P')| \leq |\rho_b(P,P'|$.

Exemple :

Dans l'exemple de référence, on trouve $\rho_a(P,P') = -.73$
et $\rho_b(P,P') = -.76$.

5 - CONCLUSION

L'objet de cette étude était d'essayer de clarifier la jungle des coefficients d'accords entre deux préordres totaux rencontrés dans la littérature et d'aider au choix d'un tel coefficient. En fait nous n'avons pas retenu tous les coefficients existants, mais par contre nous en avons introduit quatre autres (τ_2 , τ_4 , τ_γ et ρ_1). Au total nous avons étudié seize coefficients dont on trouvera en annexe 1 une présentation résumée. Le tableau de la page suivante apporte une première clarification en classant ces seize coefficients.

On constate sur ce tableau qu'il y a treize coefficients généralisant tau contre trois généralisant rho . Tous les coefficients généralisant tau s'expriment en fonction des "paramètres fondamentaux", c'est-à-dire en fonction des nombres d'accords et de désaccords de différents types (cf. 2.3 - *Partition et paramètres fondamentaux associés à une paire de préordres totaux*). Les sept coefficients de la case Gauche-Haut du tableau ont pour dénominateur commun le nombre $n(n-1)$ de couples $(x \neq y)$; leur numérateur comporte une quantité fixe $2(a-d)$, i.e. la différence entre le nombre des accords et co-accords stricts et celui des désaccords stricts, et une quan-

		GENERALISATION DE	
		TAU	RHO
NORMALISATION DE	DISSIMILARITE	τ_1 , τ_2 , τ_3 , τ_4 , τ_5 τ_a τ_F	ρ_1
	PRODUIT SCALAIRE	τ_a , e , τ_b , γ τ_γ d_a , d_b	ρ_a ρ_b

tité variable résultant des pondérations diverses attribuées aux accords larges et aux semi-accords ou désaccords. Cette structure linéaire rend la comparaison de ces coefficients particulièrement simple; en leur ajoutant tous ceux qui pourraient être construits d'une manière analogue (cf. remarque 3, paragraphe 3.3), on dispose là d'une gamme étendue de coefficients où le choix est ramené à une réflexion sur les pondérations qu'on veut attribuer aux différents types d'accord et de désaccord. Les six coefficients de la case Gauche-Bas du tableau ont tous le même numérateur $2(a-d)$, mais des dénominateurs variables obtenus en retranchant du nombre n^2 des couples

(x,y) un certain nombre de couples de semi-accords, semi-désaccords ou accords larges. Bien que ces coefficients soient encore comparables, l'effet produit par cette "pénalisation" de certains couples est beaucoup moins clair que dans le cas précédent. C'est toutefois dans ces coefficients qu'on trouve les deux seuls coefficients, parmi les treize généralisant tau étudiés ici, pour lesquels les valeurs extrêmes $+1$ ou -1 sont obtenues si et seulement si $P = P'$ $(\neq X^2)$ ou $P = P'^r$ $(\neq X^2)$, condition qu'on peut vouloir être réalisée dans certains cas (ces deux coefficients sont e et τ_b). Cette même condition est aussi vérifiée pour le seul coefficient ρ_b parmi les trois coefficients généralisant rho. On notera que parmi les seize coefficients étudiés ρ_b, τ_b et τ_γ sont les seuls pouvant être définis comme coefficients de corrélation linéaire. De fait, il n'y a a priori aucune raison pour qu'un coefficient de corrélation linéaire, mesurant la dépendance linéaire, soit pertinent comme coefficient d'accord, mesurant l'identité.

Comme nous l'avons dit dans l'introduction, notre étude a porté sur des préordres totaux arbitraires définis sur un ensemble X non structuré. Si on impose des conditions aux préordres totaux considérés, par exemple que l'un soit à deux classes, ou à l'ensemble X, par exemple qu'il soit l'ensemble des paires d'éléments d'un autre ensemble, on est dans le cadre de problèmes spécifiques : définition d'un coefficient "rang-bisérial" (cf., par exemple, Lignon, 1981), d'un coefficient d'accord entre préordonnances, etc..., problèmes sur lesquels nous reviendrons ultérieurement. Nous reviendrons aussi sur la comparaison de nos seize coefficients d'accord, d'une part pour étudier systématiquement l'ordre entre ces coefficients, d'autre part pour comparer les préordonnances qu'ils induisent sur l'ensemble des paires de préordres totaux (cf. V. Giakoumakis et B. Monjardet, 1987).

Pour terminer signalons deux autres voies de recherche intéressantes. La première consisterait dans l'approche "axiomatique" des coefficients d'accords. Il s'agit de caractériser chaque coefficient par une série de propriétés. On peut signaler à ce propos que le même problème dans le cas des dissimilarités entre préordres totaux n'a été résolu à notre connaissance que pour les cas de la distance de la différence symétrique (cf. Barthélémy, 1979) et de la distance de la norme L_1 entre les vecteurs rangs (Cook et Seiford, 1978). Plus généralement, on peut chercher à caractériser

une classe de coefficients d'accord vérifiant certaines propriétés jugées essentielles dans un certain domaine. La seconde voie de recherche consisterait à généraliser les coefficients d'accord précédents au cas de k préordres totaux; une approche simple de ce problème utilise des moyennes de coefficients ainsi définis, qui n'auront pas forcément les interprétations géométriques qu'ont les moyennes des tau de Kendall ou des rho de Spearman dans le cas des ordres totaux (cf. à ce sujet Monjardet, 1985).

6. REFERENCES

BARBUT, M., MONJARDET, B., "Ordre et classification, algèbre et combinatoire", Hachette, Paris, 1970.

BARTHELEMY, J.P., "Propriétés métriques des ensembles ordonnés. Comparaison et agrégation des relations binaires", thèse d'état, Besançon, 1979.

BARTHELEMY, J.P., "Caractérisation axiomatique de la distance de la différence symétrique entre des relations binaires", Math. et Sci. hum., 1979, 67, 85-113.

BATTEAU, P., JACQUET-LAGREZE, E., MONJARDET, B., (édit) "Analyse et agrégation des préférences", Economica, Paris, 1981.

BENZECRI, J.P. "L'analyse des données", tome 1 : la taxinomie, Dunod, Paris, 1973.

CAILLEZ, F., PAGES, J.P., "Introduction à l'analyse des données, SMASH, Paris, 1976.

CHAH. S., "Calcul des partitions optimales d'un critère d'adéquation à une préordonnance", Pub. Inst. Stat. Univ. Paris, 1984, 29, 1, 27-45.

CHANDON, J.L., BOCTOR, F.F., "Approximation d'une préordonnance par une partition", RAIRO, 1985, 19, 2, 159-184.

FALGUEROLLES (de), A., "Classification automatique : un critère et des algorithmes d'échange", in Seminaire INRIA Classification automatique et perception par ordinateur, 1977, 29-40.

GIAKOUMAKIS, V., "Coefficients d'accord entre deux préordres totaux", these de 3e cycle, Université de Paris V, 1985.

GIAKOUMAKIS, V., MONJARDET, B., "Coefficients d'accord entre deux préordres totaux II. Comparaison ordinale des coefficients, Math. et Sci. hum., 1987, 98, 69-87.

GOODMAN, L.A., KRUSKAL, W.H., "Measures of Association for Cross Classification", Springer-Verlag, New-York, 1979.

HUBERT, L.J., ARABIE, P., "Comparing Partitions" Journal of Classification, 1985, 2, 2-3, 193-218.

HUBERT, L.J., GOLLEDGE, R.G., COSTANZO, C.M., GALE, N., "Order-Dependent Measures of Correspondence for Proximity Matrices and Related Structures, in Measuring the Unmeasurable, eds, P. Nijkamp and H. Leitner, The Hague, Martinus Nijhoff, 1985.

KENDALL, M.G., "Rank Correlation Methods", 4th ed., London, Griffin, 1970.

KRUSKAL, J.B., "Multidimensionnal Scaling by Optimizing Goodness of Fit to a nonmetric hypothesis" et "Nonmetric Multidimensionnal Scaling : a Numerical Method", Psychometrika, 1964, 29, 1-27 et 115-129.

LECALVE, G., "Un indice de similarité pour des variables de type quelconque", Stat. et Anal. des Données, 1976, 01-02, 39-47.

LEMAIRE, J., "Agrégation typologique de données de préférences", Math. Sci. hum., 1977, 68, 31-50.

LERMAN, I.C., "Les bases de la classification automatique", Gauthier-Villars, Paris, 1970.

LERMAN, I.C., "Etude distributionnelle de statistiques de proximité entre structures algébriques finies du même type ; application à la classification automatique", Cahiers du BURO, 1973, 19.

LERMAN, I.C., "Classification et analyse ordinale des données", Dunod, Paris, 1981.

LIGNON, Y., "Corrélation entre deux variables dont l'une est dichotomisée", Math. Sci. hum., 1971, 76, 47-57.

LINGOES, J.C., "Indices of Configural Similarity", in Geometric Representations of Relational Data, Lingoes J.C., Roskam E.E., Borg I., edit., Mathesis Press, Michigan, 1979; 675-679.

MIRKIN, B., "Group Choice", V.H. Winston Sons, Washington, 1979.

MONJARDET, B., "Concordance et consensus d'ordres totaux, les coefficients K et W", Revue Statistique Appliquée, 1985, 23, 2, 55-85.

MONJARDET, B., LECONTE DE POLY-BARBUT, C., "Valeurs extrémales de la différence des deux coefficients de corrélation de rang ρ et τ" C.R.A.S., Paris, t. 303, Série 1, n° 10, 1986, 483-456.

SCHADER, M., "Hierarchical Analysis : Classification with Ordinal Object Dissimilarities", Metrika, 1980, 27, 127-132.

SHEPARD, R.N., "The Analysis of Proximities : Scaling with an Unknown Distance Function", Psychometrika, 1962, 27, I : 125-140,II : 219-246.

SIBSON, R., "Order Invariant Methods for Data Analysis", J. Roy. Statistical Soc., B., 1972, 34, 311-349.

SOKAL, R.R., SNEATH, P.H.A., "Principles of Numerical Taxonomy", W.H. Freeman and Company, San Francisco, 1963.

SOMMERS, R.H., "A New Asymetric Measure of Association for Ordinal Variables", Amer. Soc. Rev., 1962, 27, 799-811.

SPEARMAN, C., "A Footrule for Measuring Correlation", Brit. J. Psychol., 1906, 2, 89-108.

VEGA (de la), C.F., "Techniques de classification automatique utilisant un indice de ressemblance", Revue Française de Sociologie, 1967, 8, 506-520.

WILSON, T.P., "Measures of Association for Bivariate Hypothesis", in H.M. Blalock (ed.), Measurement in the Social Sciences, Chicago, Aldine, 1974.

ANNEXE I - CODAGE D'UN PREORDRE TOTAL P = O+I

Codages dans $\mathbb{R}^{n^2}$

Codage α

$\alpha(x,y) = +1 \Leftrightarrow (x,y) \in P$

$\alpha(x,y) = 0 \Leftrightarrow (x,y) \notin P$

Codage γ

$\gamma(x,y) = +1 \Leftrightarrow (x,y) \in P$ et $x \neq y$

$\gamma(x,y) = -1 \Leftrightarrow (x,y) \notin P$ et $x \neq y$

$\gamma(x,y) = 0 \Leftrightarrow x = y$

Codage β

$\beta(x,y) = +1 \Leftrightarrow (x,y) \in P$ et $(y,x) \notin P \Leftrightarrow (x,y) \in O$

$\beta(x,y) = -1 \Leftrightarrow (x,y) \notin P$ et $(y,x) \in P \Leftrightarrow (x,y) \in O^r$

$\beta(x,y) = 0 \Leftrightarrow (x,y) \in P$ et $(y,x) \in P \Leftrightarrow (x,y) \in I$

Codages dans $\mathbb{R}^n$

$P = C_1 < C_2 \ldots < C_t$ $\qquad n_\ell = |C_\ell|$, $\ell = 1, \ldots t$

$x_1 \ldots x_i \ldots x_n$ est un ordre total sur X ; on donne le codage de l'élément x_i appartenant à la classe C_k :

Codage "rang" r^-

$$r^-(x_i) = |\{y \in X : y P x_i\}| = \sum_{\ell=1}^{k} n_\ell$$

Codage "rang moyen" r

$$r(x_i) = r^-(x_i) - \frac{n_k - 1}{2} = \sum_{\ell<k} n_\ell + \frac{n_k + 1}{2}$$

Codage "score"

$$s(x_i) = |\{y \in X : x_i P y\}| - |\{y \in X : y P x_i\}| = n + n_k - 2r^-(x_i) = n + 1 - 2r(x_i)$$

ANNEXE II - LISTE DES SEIZE COEFFICIENTS D'ACCORD ENTRE DEUX PREORDRES TOTAUX P ET P'

	FORMULE DE DEFINITION	FORMULE DE CALCUL	AUTEURS	§
τ_1	$1 - \frac{2\delta_H(\vec{\beta},\vec{\beta}')}{n(n-1)}$	$\frac{2(a-d) + 2(c^*-b)}{n(n-1)}$	LINGOES (m)	3.3.1
τ_2	$1 - \frac{\delta_H(\vec{\beta},\vec{\beta}')+\delta_\Delta(P,P')}{n(n-1)}$	$\frac{2(a-d) + (2c^*-b)}{n(n-1)}$		3.3.2
τ_3	$1 - \frac{2\delta_\Delta(P,P')}{n(n-1)}$	$\frac{2(a-d) + 2c^*}{n(n-1)}$	De la VEGA, LERMAN SOKAL et MICHENER	3.3.3
τ_4	$1 - \frac{\delta^2_E(\vec{\beta},\vec{\beta}')}{2n(n-1)}$	$\frac{2(a-d) + 2c^* + b}{n(n-1)}$		3.3.4
τ_5	$1 - \frac{4\lvert O \cap O'^{\overline{x}}\rvert}{n(n-1)}$	$\frac{2(a-d) + 2(c^*+b)}{n(n-1)}$	LINGOES (m*) LERMAN	3.3.5
τ_F	$1 - \frac{2\lvert \overset{o}{P} \Delta O'\rvert}{n(n-1)}$	$\frac{2(a-d) - 2c^*}{n(n-1)}$	FALGUEROLLES	3.3.7
τ_a	$\frac{\vec{\beta}.\vec{\beta}'}{n(n-1)}$	$\frac{2(a-d)}{n(n-1)}$	KENDALL	3.3.6 4.2.1
e	$\frac{\vec{\beta}.\vec{\beta}'}{n^2 - \lvert I \cap I'\rvert}$	$\frac{a-d}{a+b+d}$	WILSON	4.2.1
d_a	$\frac{\vec{\beta}.\vec{\beta}'}{n^2 - \lvert I\rvert}$	$\frac{a-d}{a+b_2+d}$	SOMERS	4.2.1
d_b	$\frac{\vec{\beta}.\vec{\beta}'}{n^2 - \lvert I'\rvert}$	$\frac{a-d}{a+b_1+d}$	SOMERS	4.2.1
τ_b	$\text{corr}\ (\vec{\beta},\vec{\beta}');\ \frac{\vec{\beta}.\vec{\beta}'}{2(\lvert O\rvert.\lvert O'\rvert)^{\frac{1}{2}}}$	$\frac{a-d}{[(a+b_1+d)(a+b_2+d)]^{\frac{1}{2}}}$	KENDALL	4.2.1 4.3
γ	$\frac{\vec{\beta}.\vec{\beta}'}{n^2 - \lvert I \cup I'\rvert}$	$\frac{a-d}{a+d}$	GOODMAN et KRUSKAL	4.2.1
τ_γ	$\text{corr}\ (\vec{\gamma},\vec{\gamma}')$			4.3
ρ_1	$1 - \frac{2\delta^2_E(\vec{r},\vec{r}')}{\max \delta^2_E(\vec{r},\vec{r}')}$	$1 - \frac{6\sum_i (r_i - r'_i)^2}{n^3 - n}$		3.4
ρ_a	$\frac{12\ \text{Cov}\ (\vec{r},\vec{r}')}{n^2 - 1}$		KENDALL	4.3
ρ_b	$\text{corr}\ (\vec{r},\vec{r}')$		KENDALL	4.3

$P = O + I$ $\quad a = \lvert O \cap O'\rvert$ $\quad d = \lvert O \cap O'^{\overline{x}}\rvert$ $\quad c^* = \frac{\lvert I \cap I'\rvert}{2} - n$

$P' = O' + I'$ $\quad b_1 = \lvert I \cap O'\rvert$ $\quad b_2 = \lvert I' \cap O\rvert$ $\quad b = b_1 + b_2$

$\overset{o}{P} = P - \{(x,x)\ ,\ x \in X\}$

ANNEXE III - CARACTERISATION DES COUPLES (P,P') REALISANT LES VALEURS EXTREMALES DU COEFFICIENT D'ACCORD

COEFFICIENT	VALEUR +1	VALEUR -1
τ_1	$P = P'$	$P \cap P'^r$ ordre total
τ_2	$P = P'$	$P = P'^r$ ordre total
τ_3	$P = P'$	$P = P'^r$ ordre total
τ_4	$P = P'$	$P = P'^r$ ordre total
τ_5	$P \cap P'$ préordre total	$p = P'^r$ ordre total
τ_a	$P = P'$ ordre total	$P = P'^r$ ordre total
τ_F	$P = P'$ ordre total	$P = P'^r$
e	$P = P' \neq X^2$	$P = P'^r \neq X^2$
d_a	$P' \subseteq P \neq X^2$	$P' \subseteq P^r \neq X^2$
d_b	$P \subseteq P' \neq X^2$	$P \subseteq P'^r \neq X^2$
τ_b	$P = P' \neq X^2$	$P = P'^r \neq X^2$
γ	$P \cap P'$ préordre total $\neq X^2$	$P \cap P'^r$ préordre total $\neq X^2$
τ_γ	$P = P' \neq X^2$	$P = P'^r$ ordre total
ρ_1	$P = P'$	$P = P'^r$ ordre total
ρ_a	$P = P'$ ordre total	$P = P'^r$ ordre total
ρ_b	$P = P' \neq X^2$	$P = P'^r \neq X^2$