REVUE DE STATISTIQUE APPLIQUÉE

LUU-MAU-THANH
SULLY LEDERMANN

Étude d'empreintes digitales. Mesure d'association par l'analyse canonique

Revue de statistique appliquée, tome 11, n° 4 (1963), p. 77-84
<http://www.numdam.org/item?id=RSA_1963__11_4_77_0>

ETUDE D'EMPREINTES DIGITALES
MESURE D'ASSOCIATION PAR L'ANALYSE CANONIQUE

LUU-MAU-THANH et Sully LEDERMANN (1)
Institut National d'Etudes Démographiques

L'objet de la présente étude est limité à la mesure du degré de ressemblance entre les doigts homologues des deux mains : pouce avec pouce, index avec index, etc.

La ressemblance des doigts s'exprime ici à l'aide de celle des empreintes digitales. Les dessins de ces empreintes ont une forme complexe. En première approximation, ils ont été classés en quatre catégories : arc (A), boucle radiale (Br), boucle cubitale (Bu) et tourbillon (T). Les deux catégories de boucles se différentient par leur orientation : vers le pouce (boucles radiales), vers l'auriculaire (boucles cubitales).

Les données numériques de travail sont des tableaux de contingence donnant les fréquences d'association des diverses catégories (cf. tableau 1), par paire de doigts (tableaux 2 à 6) et pour l'ensemble des doigts appariés (tableau 7).

Tableau 1

Tableau de contingence de 2 caractères A et B

	B_1	B_2	———	B_j	——	B_s	$\sum B_j$
A_1	n_{11}	n_{22}	———	n_{1j}	——	n_{1s}	n_{10}
A_2	n_{21}	n_{22}	———	n_{2j}	——	n_{2s}	n_{20}
A_i	n_{i1}	n_{i2}	- - - - - - - - - -	n_{ij}	- - -	n_{is}	n_{i0}
A_r	n_{r1}	n_{r2}	- - - - - - - - - -	n_{rj}	- - -	n_{rs}	n_{r0}
$\sum A_i$	n_{01}	n_{02}	- - - - - - - - - - -	n_{0j}		n_{0s}	n_{00}

* * *

(1) Les données de base ont été élaborées au Centre de recherches anthropologiques (Directeur : Docteur Gessain). L'étude a été faite par M. Luu-Mau-Thanh, chargé de mission à l'INED et statisticien au Centre de recherches anthropologiques, en collaboration avec M. Sully Ledermann de l'Institut national d'études démographiques.

Il n'y a pas qu'une solution au problème de la mesure de la ressemblance entre unités, objets ou individus, associés deux par deux.

a) Si les critères retenus, ici 4 formes, ne constituent pas un ordre hiérarchique supposant un continuum uni-dimensionnel latent, la mesure du degré de ressemblance entre les doigts ne peut être qu'une mesure d'association. On peut utiliser les coefficients de contingence de Pearson, de Tschuprow ou de Cramer, basés sur la valeur du χ^2 du tableau de contingence. On peut songer également aux coefficients proposés par Goodman-Kruskal conçus autour de la probabilité de prévoir correctement une des variables, l'autre étant donnée.

b) Si les critères se prêtent à un ordre hiérarchique, le problème peut se traiter autrement en recourant aux divers coefficients de corrélation de rangs. Aller plus loin dans cette voie suppose l'existence d'une métrique, c'est-à-dire d'un continuum uni-dimensionnel latent, dont la métrique n'est pas saisissable directement.

Une solution est d'imposer cette métrique et on est alors ramené au calcul de coefficient de corrélation d'une population bi-variate. La difficulté est de décider quelles valeurs prendre pour cette métrique, quelle codification numérique choisir, d'autant plus qu'on cherche une mesure satisfaisante de similitude et non pas seulement un critère permettant de tester l'indépendance.

La métrique la plus simple, en l'absence totale d'information, est une suite de nombres entiers : 1, 2, 3 ou - 1,0, + 1 dans le cas de 3 catégories ordonnées, par exemple. C'est la suite - 1,0, + 1 qui avait été retenue par Turpin et Schutzenberger en 1949[1] dans l'analyse d'un matériel analogue au nôtre, mais où les critères de forme retenus n'étaient que : arc, boucle (sans distinction de direction) et tourbillon.

Une autre métrique possible est liée à une hypothèse de normalité du continuum latent des critères.

Une autre codification numérique enfin est constituée par les valeurs numériques telles que le coefficient de corrélation obtenu soit maximum. C'est ce que nous avons essayé ici sur 4 formes d'empreintes digitales. Un avantage de cette dernière approche est de donner l'ordre des critères dans le continuum uni-dimensionnel latent, lorsque cet ordre lui-même n'est pas évident,... si le continuum uni-dimensionnel existe.

* * *

Le principe et la technique de la codification canonique sont décrits par Kendall et Stuart[2]. Nous en rappellerons seulement l'essentiel.

Les tableaux 2 à 6 donnent les fréquences d'association des 4 formes pour les doigts appariés : pouce avec pouce, index avec index, etc.

Pour certaines utilisations ultérieures, nous avons cherché plutôt une codification numérique moyenne des formes, relative à l'ensemble des doigts appariés. Pour l'obtenir, on pourrait effectuer 5 analyses canoniques et prendre la moyenne des valeurs numériques obtenues. Mais

---

(1) Turpin (R.) et Schützenberger (1949). L'étude des dermatoglyphes. Sem. Hôp. Paris, 14 août 1949.

(2) Kendall (M.G.) and Stuart (A.). The advanced Theory of Statistics. London 1961, Charles Griffin & Company limited, vol. II, pp. 566-578.

Tableau 2

Pouces

		droit				
	Figures	A	Br	Bu	T	Σ
gauche	A	86	2	39	4	131
	Br	3	2	8	4	17
	Bu	128	9	2539	294	2970
	T	21	9	1008	1629	2667
	Σ	238	22	3594	1931	5785

Tableau 3

Index•

		droit				
	Figures	A	Br	Bu	T	Σ
gauche	A	359	168	206	16	749
	Br	127	512	393	291	1323
	Bu	197	307	989	216	1709
	T	25	204	494	1410	2133
	Σ	708	1191	2082	1933	5914

Tableau 4

majeurs

		droit				
	Figures	A	Br	Bu	T	Σ
gauche	A	228	16	177	4	425
	Br	9	12	81	20	122
	Bu	300	104	3483	341	4228
	T	1	3	476	750	1230
	Σ	538	135	4217	1115	6005

Tableau 5

Annulaires

		droit				
	Figures	A	Br	Bu	T	Σ
gauche	A	71	2	48	1	122
	Br	5	4	41	9	59
	Bu	72	18	2557	249	2896
	T	7	6	1082	1848	2943
	Σ	155	30	3728	2107	6020

Tableau 6

Auriculaire

		droit				
	Figures	A	Br	Bu	T	Σ
gauche	A	61	1	25		87
	Br	1	1	11	3	16
	Bu	45	2	4687	188	4922
	T	1	2	503	496	1002
	Σ	108	6	5226	687	6027

les calculs sont lourds et il a paru préférable de n'effectuer qu'une seule analyse canonique portant sur le tableau 7 dont les cases contiennent les sommes des fréquences des cases correspondantes dans les tableaux 2 à 6. Ce tableau des sommes serait assez difficile à traiter si on voulait aborder des fluctuations d'échantillonnage, par exemple, mais la question ne se pose pas. Son seul intérêt est de permettre d'obtenir une codification moyenne par une seule analyse canonique au lieu d'en obtenir une, probablement peu différente, par 5 analyses.

Soit $x_1\, x_2\, x_3\, x_4$ la codification numérique cherchée pour les figures : A, Br, Bu et T de la main gauche : $y_1\, y_2\, y_3\, y_4$ cette codification des mêmes figures pour la main droite. En notations matricielles, posons (l'indice "prime" indiquant une matrice transposée) :

$$X' = (x_1\, x_2\, x_3\, x_4) \qquad (1)$$

$$Y' = (y_1\, y_2\, y_3\, y_4) \qquad (2)$$

avec : $E(X) = E(Y) = 0$ (3)

Soit N_G et N_D les matrices diagonales dont les éléments sont les fréquences marginales du tableau de contingence (mains gauche et droite) et N ce tableau lui-même (sans les fréquences marginales). (Tableau 7).

Tableau 7

Tableau des fréquences globales (matrice N)

		main droite				
		A (1)	Br (2)	Bu (3)	T (4)	Σ
main gauche	(1) A	805	189	495	25	1 514
	(2) Br	145	531	534	327	1 537
	(3) Bu	742	440	14 255	1 288	16 725
	(4) T	55	224	3 563	6 133	9 975
	Σ	1 747	1 384	18 847	7 773	29 751

Le coefficient r de corrélation a pour valeur :

$$r = \frac{1}{n_{oo}} X' NY \tag{4}$$

On cherche à maximiser r avec les conditions supplémentaires de variance :

$$V(X) = 1 = \frac{1}{n_{oo}} X' N_G X \tag{5}$$

$$V(Y) = 1 = \frac{1}{n_{oo}} Y' N_D Y \tag{6}$$

La valeur cherchée de r est la racine la plus grande du polynôme en r^2, développement du déterminant (tableau 8) :

$$|r^2 I - N_G^{-1}. N. N_D^{-1}. N'| = 0 \tag{7}$$

Tableau 8

Matrice $(N_p^{-1} \times N \times N_q^{-1} \times N')$

270 691 522	101 984 816	515 541 881	111 781 661
100 458 702	159 174 022	447 936 235	292 431 013
046 668 491	041 164 602	684 620 606	227 546 272
016 966 159	045 059 295	381 524 957	556 449 569

C'est un polynôme du 4ème degré en r^2 :

$$r^8 - 1,670\,937\ r^6 + 0,802\,908\ r^4 - 0,138\,986\ r^2 + 0,007\,015 = 0$$

dont les 4 racines en r^2 sont : 1 ; 0,354 55 ; 0,230 515 ; 0,085 88.

La racine $r = 1$ correspond à la solution $x_1 = x_2 = \ldots = y_3 = y_4$. C'est une solution exclue, qui donnerait évidemment une corrélation égale à 1.

Nous retenons la corrélation la plus élevée $r^2 = 0,35455$ d'où $r = 0,595\ 35$.

Les valeurs de X et Y correspondantes sont les solutions du système :

$$\left\{\begin{array}{c|c} -rN_6 & N \\ \hline N' & -rN_0 \end{array}\right\} \left\{\begin{array}{c} X \\ \hline Y \end{array}\right\} = 0 \qquad (8)$$

Elles sont définies à un facteur de proportionnalité près. Les solutions correspondant aux conditions (5) et (6) sont :

	Main gauche	Main droite
A	- 2,148	- 2,045
Br	- 0,403	- 0,690
Bu	- 0,555	- 0,400
T	+ 1,319	+ 1,554

Telle est la codification numérique moyenne donnée par la "canonisation" du tableau 7.

1/ La codification numérique précédente est seulement une série de valeurs entre une infinité d'autres, auxquelles on passe par un changement de variable linéaire. Toutes ces codifications numériques donnent la même corrélation canonique. Elles sont équivalentes.

On peut rapprocher ainsi la codification présente de celle utilisée par Turpin et Schützenberger - 1, 0, + 1 pour les figures Arcs, Boucles (sans distinction), Tourbillons. Les deux codifications sont en relation sensiblement linéaires et pratiquement équivalentes, sous réserve de la confusion des boucles radiales et cubitales, qui paraît ici justifiée à posteriori par la proximité de nos valeurs.

2/ En adoptant cette codification moyenne, les coefficients de corrélation des paires de doigts : pouce - pouce, index - index, etc. sont, après calcul selon les fréquences figurant dans les tableaux 2 à 6 :

pouces	: 0,563
index	: 0,588
majeurs	: 0,576
annulaires	: 0,589
auriculaires	: 0,555

Les valeurs des coefficients sont très voisines. Aucune paire de doigts homologues ne se fait remarquer par une similitude ou une dissemblance plus particulièrement marquée que les autres.

Il est intéressant de rapprocher ces résultats de l'information apportée par la probabilité de trouver deux doigts homologues identiques, abstraction faite de la figure à l'origine de l'identité. Cette probabilité obtient facilement en calculant le rapport de la somme des termes de la diagonale des tableaux 2 à 6, au total des observations de chaque tableau. On trouve :

	Probabilité		A		Br		Bu		T
Pouces.....	0,74	=	0,01	+	0	+	0,45	+	0,28
Index.....	0,55	=	0,06	+	0,09	+	0,17	+	0,24
Majeurs.....	0,74	=	0,04	+	0	+	0,58	+	0,12
Annulaires.....	0,74	=	0,01	+	0	+	0,42	+	0,31
Auriculaires.....	0,87	=	0,01	+	0	+	0,78	+	0,08

Les doigts apparaissent ici plus différentiés que ne le suggèrent les corrélations. Mais on doit remarquer que la probabilité faible 0,55 de similitude des index, par exemple, est essentiellement due à la différentiation des boucles entre radiales cubitales, dont l'association est bien plus fréquente pour les index que pour les autres doigts (tableau 3). Dans la comparaison des doigts, la codification canonique tient compte de la proximité des deux formes de boucles.

En sens contraire, la corrélation canonique 0,55 pour les auriculaires subit, par rapport à la probabilité 0,87, l'influence de la distance relativement grande existant dans l'association tourbillon et boucle cubitale, fréquente pour les auriculaires et qui prend d'autant plus d'importance que les autres associations sont peu représentées (tableau 6).

La probabilité de similitude est d'une interprétation plus facile, mais le coefficient de corrélation canonique tient compte de diverses associations et, en plus, des distances respectives des figures au sein d'une certaine structure. L'information qu'il apporte est donc plus complète.

3/ Les doigts non homologues présentent des corrélations moins élevées que celles des doigts homologues (cf. tableau 9). Nous ne nous étendrons pas sur leurs différences qui feront l'objet d'une étude particulière.

CONCLUSION

La transposition de l'analyse canonique dans la mesure du degré d'association entre individus ou objets caractérisés par une série de critères qualitatifs peut apporter une aide certaine en proposant une métrique pour ces critères, sans considération à priori sur leur ordre respectif dans un continuum latent uni-dimensionnel. La "codification canonique" fournit à la fois l'ordre des critères, s'il n'est pas évident, et une idée de leurs distances respectives.

Mais son application suppose l'existence de ce continuum latent. L'analyse canonique n'en apporte pas la preuve.

Les critères intervenant dans notre étude sont des formes d'empreintes digitales : arc, boucles radiale et cubitale, tourbillon. Les boucles radiale et cubitale ne diffèrent que par leur orientation.

La "codification canonique" obtenue conduit à considérer comme très voisines les boucles radiale et cubitale, par rapport aux distances qui les séparent des formes arc et tourbillon. Elles le sont en effet : ce sont les mêmes boucles, à l'orientation près. Mais ce dernier détail soulève le problème de fond. Dans quel continuum uni-dimensionnel, ces diverses distances prennent-elles un sens ? La réponse à cette question relève non plus de la statistique mathématique, mais de la génétique et de l'anthropologie.

Tableau 9

Matrice de corrélations entre les différents doigts[1] (en millièmes)

		Main gauche					Main droite				
		1	2	3	4	5	6	7	8	9	10
Main gauche	1	-	338	303	325	254	*563*	348	293	276	203
	2	338	-	490	449	334	324	*588*	473	421	288
	3	303	490	-	462	341	284	493	*576*	446	259
	4	325	449	462	-	401	263	446	458	*589*	318
	5	254	334	341	401	-	214	346	333	419	*555*
Main droite	6	*563*	324	284	263	214	-	360	294	288	230
	7	348	*588*	493	446	346	360	-	497	431	297
	8	293	473	*576*	458	333	294	497	-	503	317
	9	276	421	446	*589*	419	288	431	503	-	416
	10	203	288	259	318	*555*	230	297	317	416	-

(1) Les chiffres en italique sont les corrélations des doigts homologues entre eux : pouce avec pouce (1 et 6), index avec index (2 et 7), etc.