Un exemple d'application de l'analyse factorielle des correspondances : étude de neuf lots de souris blanches

L. BELLIER,
Ecologiste de l'O.R.S.T.O.M., Abidjan (Côte d'Ivoire).

RÉSUMÉ

Présentation d'une nouvelle méthode d'analyse statistique d'emploi très général permettant la discrimination entre groupes.

ABSTRACT

The author explains, on a simple example, a new, efficient method to analyse groups.

L'Analyse Factorielle des Correspondances, mise au point par M. le Professeur BENZECRI (Laboratoire de Statistique Mathématique de l'Université de Paris), est un outil statistique mathématique relativement récent. Elle est susceptible d'aider le chercheur à résoudre de nombreux problèmes de discrimination entre lots, en particulier lorsque ceux-ci sont constitués d'individus sur lesquels on peut prendre un grand nombre de mesures. Si N est le nombre de mesures, on peut imaginer que chaque individu est caractérisé, dans un espace dual de N dimensions, par ses N mesures, ou coordonnées. L'ensemble des points correspondant à tous les individus d'un même lot constitue un nuage de points représentatif de la population étudiée. Deux populations différentes auront leurs représentations dans cet espace à N dimensions plus ou moins disjointes. La difficulté réside dans la possibilité de visualisation de cet espace à N dimensions. Or les N mesures — nous dirons *mesures primaires* — qui ont servi à définir cet espace, sont corrélées entre elles ; on conçoit qu'elles soient sous la dépendance d'un petit nombre de *facteurs* naturels, plus ou moins indépendants entre eux, qui régissent la forme. (Ces facteurs pourront être par exemple l'âge, le sexe, la souche de l'animal...) Dans l'espace à N dimensions des *mesures primaires* le nuage des points représentant les individus ne se présentera donc pas comme une sphère également étendue dans toutes ses directions, il aura des directions principales d'allongement en nombre approximativement égal à celui des *facteurs naturels :* disons, pour prendre une image visible, que c'est comme si, N étant 3, on avait dans l'espace usuel (3 dimensions) une galette (2 directions d'allongement) ou un cigare (1 direction d'allongement) (cf. fig. 1). Ces directions principales d'allongement, déterminées par le calcul statistique sur ordinateur sont, dans l'espace des N mesures primaires, des axes nouveaux appelés *axes factoriels*, les coordonnées sur ces axes étant les *facteurs extraits* ainsi appelés par opposition aux *facteurs naturels* plus ou moins inaccessibles à la mesure, mais que justement l'on reconnaît souvent dans les *facteurs extraits* qui ne sont mathématiquement parlant que des combinaisons linéaires des *mesures primaires ;* les coefficients de ces combinaisons résultant du calcul statistique.

Pratiquement dans une étude craniométrique où le nombre N des *mesures primaires* est d'une vingtaine, on a une vision claire des proximités relatives des individus et des groupes en se bornant à considérer un nombre n = 2, 3 ou 4, de *facteurs extraits*. Le gain est donc indéniable.

On a maintenant affaire à un petit nombre de variables numériques, les n *facteurs extraits*, qui, de plus, de par leur construction même (le calcul statistique) sont non corrélées entre elles. (Ce qui explique d'ailleurs que malgré leur petit nombre, elles donnent une bonne idée du tout.)

Ceci correspond à la théorie générale de l'analyse factorielle classique, qu'elle soit menée par rotations successives ou en étudiant les composantes principales. L'analyse des correspondances utilisée ici dérive directement de cette dernière méthode, mais elle doit toutefois son efficacité grandement accrue et son universalité d'emploi à trois éléments nouveaux.

Tout d'abord, elle ne traite pas des données brutes, mais des profils. Prenons un exemple trivial où N = 2 : un animal caractérisé par son poids x et sa taille y. On peut associer à un individu le point (x, y) d'abscisse x et d'ordonnée y ; mais on peut aussi lui associer le point $\left(\frac{x}{x+y}, \frac{y}{x+y}\right)$ (ayant pour coordonnées des rapports, dont la somme est 1) ce point étant affecté de la masse (x + y) pour rappeler l'importance de l'individu : c'est en gros ce que fait l'analyse des correspondances. L'importance de cette transformation est essentielle : elle permet de mêler sans précautions particulières des individus différant grandement par l'âge et la taille (cf. fig. 2).

Ensuite la distance utilisée n'est pas la distance euclidienne classique, elle a été choisie d'après des

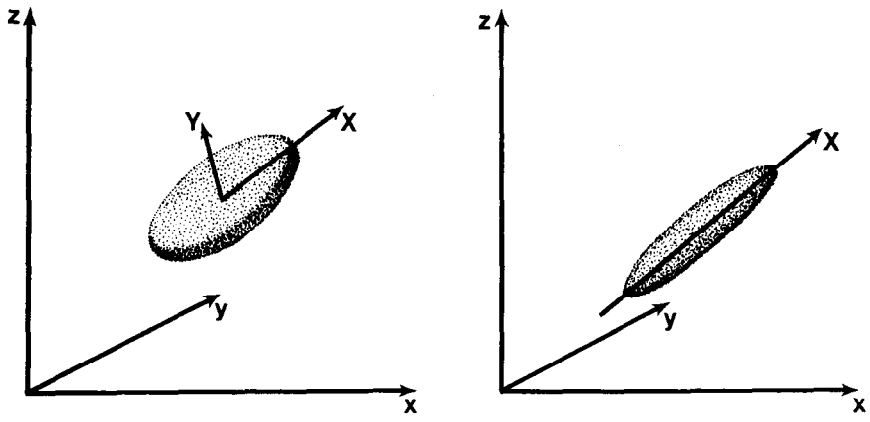

Fig. 1. — Exemples de nuages de points dans l'espace usuel à 3 dimensions : à droite, un cigare (une seule direction principale d'allongement : axe X) ; à gauche, une galette (deux directions d'étalement : X d'abord, puis Y).

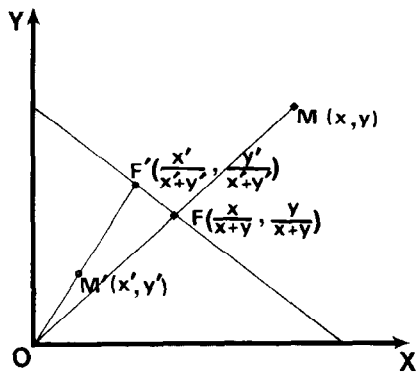

Fig. 2. — Représentation de deux individus par leurs mesures brutes (points M et M') et par leurs formes (points F et F') : il devient possible d'étudier la dispersion des formes indépendamment des inégalités de tailles.

considérations statistiques. Ici encore, il vaut mieux prendre un exemple : plaçons-nous dans un espace à deux dimensions, c'est-à-dire dans le plan ! Traçons un disque représentant un essaim de pointe. Si on dit que la distance d'un point (x, y) à l'origine est donné par la formule usuelle :

$$d = \sqrt{x^2 + y^2}$$

il n'y a pas de direction principale ; c'est comme on l'a dit, un disque. Si on pose :

$$d' = \sqrt{100\,x^2 + y^2}$$

le nuage est très allongé suivant l'axe ox (10 fois plus que suivant oy). Si au contraire on pose :

$$d'' = \sqrt{x^2 + 100\,y^2}$$

c'est oy qui est l'axe principal d'allongement. Cet exemple est troublant : il fait voir que la notion évidente de direction principale d'allongement n'est pas simple, univoque : elle dépend de ce qu'on appelle distance. On conçoit donc que les résultats de l'analyse factorielle dépendent grandement du choix judicieux d'une formule de distance. La gît le deuxième avantage annoncé.

Enfin, troisième avantage : les calculs font jouer un rôle symétrique aux individus (ici des crânes) d'une part et aux mesures de l'autre : il est donc possible de figurer sur un même graphique avec les individus l'ensemble des mesures, une classe d'individu se plaçant, en gros, autour des mesures qui sont chez elles relativement les plus développées.

Ainsi donc, en plus de la facilité de visualisation d'un espace orthogonal de dimensions réduites qu'apporte l'analyse en composantes principales, l'Analyse Factorielle des Correspondances permet de :

— traiter toutes sortes de données,

— pouvoir représenter dans le même espace les sujets et les objets concernés par l'étude.

Ainsi deux populations différant par deux caractères auront leurs représentations associées différemment à la représentation de ces caractères.

Enfin — et ce n'est pas là une qualité mineure — la méthode met en quelque sorte à l'index l'individu exceptionnel, celui qui rompt l'harmonie du lot soit simplement parce qu'il est exceptionnel et qu'il ne faut pas en tenir compte, soit parce que les mesures qu'on lui rapporte sont erronées.

L'exemple que l'on donne ci-dessous est assez typique. Dans le cadre de notre collaboration avec le laboratoire de Psychobiologie et Psychopharmacie que dirige M. le Professeur Bovet, à Rome, et qui dépend du Centre National de Recherches Italien, neuf lots de crânes de souris blanches — animaux que nous n'utilisons jamais dans nos propres recherches — nous ont été envoyés pour études biométriques. Ces lots étaient désignés par quelques initiales désignant la souche, suivies d'un nombre et d'une lettre M ou F indiquant le nombre d'individus du lot et leur sexe, Mâle ou Femelle.

Nous ignorions tout de l'antécédent des lots. Ils pouvaient avoir été soumis à des traitements différents (diète spéciale, par exemple). Chaque lot devait être composé de spécimens du même âge. Cet âge pouvait être le même pour l'ensemble des lots.

Il fallait donc pouvoir préciser, au terme de notre étude :

— si les lots étaient homogènes ou non,

— si les lots de même souche pouvaient ou non être confondus, ou devaient être considérés comme distincts.

Le matériel.

Les neuf lots sont désignés par un numéro.

La correspondance avec leur appellation d'origine est la suivante :

			Représentation graphique
Lot nº 1	SEC	24 M	rond noir
Lot nº 2	SEC	5 M	triangle noir
Lot nº 3	SEC	10 F	étoile noire
Lot nº 4	S x C	14 M	étoile blanche dans rond noir
Lot nº 5	BALB	15 M	étoile blanche
Lot nº 6	CDR	11 M	maison renversée
Lot nº 7	C 57 BL	15 M	rond blanc
Lot nº 8	C 57 BL	8 F	carré blanc
Lot nº 9	C 57 BL	12 F	triangle blanc

Le traitement des crânes s'est fait selon les procédés classiques du laboratoire :

Numérotation des crânes,

Nettoyage initial par des Dermestes,

Nettoyage final à la pince après traitement chimique,

Blanchiment à l'eau oxygénée faible.

Les crânes ont ensuite été mesurés, à raison de 17 caractères pour chaque crâne. Le détail des mensurations est inutile ici, à l'exception de la première mesure qui correspond à la longueur totale du crâne : il s'agit de la longueur condylo basale-incisive, indiquée par la suite sous l'abréviation Cb.

Il y avait au total :

I = 114 individus mesurés

pour J = 17 caractères par individu.

La matrice des données initiales est donc un tableau complet de 114 × 17.

Le couple (i, j) $i \in I$; $j \in J$ constitue donc la J^{eme} mesure effectuée sur le i^{eme} individu.

Les résultats de l'analyse.

Seules les trois premières valeurs propres ont été extraites pour cette analyse. Elles sont obtenues par ordre décroissant. Les trois facteurs correspondant extraient 66% de l'inertie totale du muage. Comme d'habitude dans les études craniométriques, le premier facteur est de loin le plus important puisqu'il extrait à lui seul plus de 43% de l'inertie totale.

Un seul graphique est donné en annexe : celui qui correspond à la projection du nuage sur le plan formé par les 1^{er} et 2^{e} facteurs. Les deux autres projections sont un peu moins typiques et leur étude ne pourrait qu'alourdir l'exposé.

A partir de ce graphique des remarques de deux types peuvent être faites :

— les premières concernent l'homogénéité ou l'hétérogénéité d'un lot. Celle-ci peut être due :

soit à la souche elle-même,

soit à une erreur dans la formation du lot (classes d'âges mélangées, par exemple),

soit surtout à des mesures erronées.

De ce type de remarques, on peut tirer quelques conclusions sur la comparaison de lots entre eux.

— Les remarques du deuxième ordre concerne les relations qui existent entre lots. La représentation graphique de deux lots peut être :

— confondue, et l'on ne peut conclure qu'à l'identité des deux lots,

— proche ce qui signifie qu'il y a une différence notable entre les deux lots considérés mais qu'ils peuvent dériver l'un de l'autre,

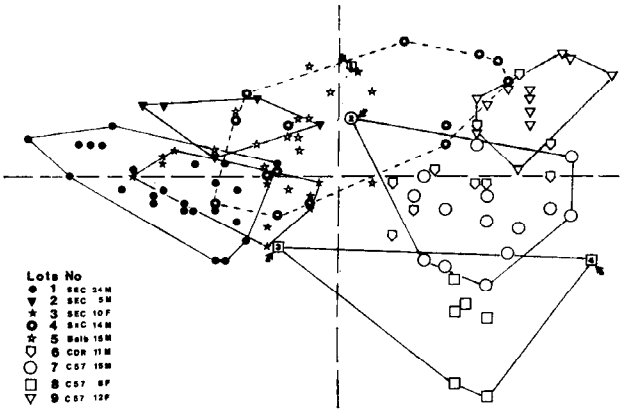

Fig. 3.

— sans relation, et dans ce cas les lots doivent être considérés comme absolument distincts.

Cette absence de relations entre lots d'une même souche peut provenir soit du choix des individus lors de la formation des lots (individus de classe d'âge ou de sexe différents) soit d'un traitement préalable connu ou inconnu.

A *Homogénéité.*

Les lots 1, 2 et 3, tous trois de la souche SEC sont remarquablement homogènes sur les trois plans de projection.

Le lot 4, souche S x C est par contre remarquablement hétérogène. Il peut, sur les trois plans de projection, être scindé en deux groupes : l'un de 6 individus, situé dans le premier quadrant du graphique, l'autre de huit spécimens, proche des lots SEC.

Les mensurations, vérifiées, ne permettent de déceler aucune erreur de mesure.

Le lot 5, souche BALB est également homogène.

Le lot 6, souche CDR, est légèrement hétérogène, et cela est dû essentiellement au spécimen marqué 1 et signalé par une flèche. La vérification des mesures montre que celles-ci ont été correctement effectuées.

Le lot 7, souche C 57 est relativement dispersé. Un rongeur s'écarte plus du lot que les autres — il est indiqué par une flèche et porte le numéro 2. L'examen des 17 mesures montre que deux d'entre elles sont erronées, expliquant ainsi l'écart du point par rapport à l'ensemble des projections du lot considéré. De même le lot 8, souche C 57 également, doit sa dispersion à deux points flechés sur le graphique et numérotés 3 et 4. Le crâne correspondant au numéro 3 a été mal mesuré : deux mensurations sont erronées, par contre l'examen des mensurations du rongeur représenté par le numéro 4 ne montre qu'une petite inexactitude portant sur une seule mensuration, ce qui n'explique pas sa position excentrique.

Enfin le dernier lot (numéro 9) de la souche C 57 est remarquablement homogène sur le graphique présenté, mais est légèrement plus hétérogène sur les

autres graphiques — une erreur de mesure a ainsi pu être décelée.

B *Relations entre lots.*

Il est plus délicat, ici, de traiter des relations entre lots, compte tenu du fait que nous ignorons tout des antécédents de ces lots et que même la signification des symboles nous échappe.

A titre d'exemple nous pouvons nous arrêter sur les deux groupes de trois lots désignés par les mêmes initiales.

En ce qui concerne la souche SEC, les lots 1 et 3 sont proches l'un de l'autre et le petit écart qui existe entre eux peut simplement traduire un léger dimorphisme sexuel.

Par contre le lot 2 est assez nettement séparé du lot 1 alors que dans les deux cas, les individus sont du même sexe.

La souche C 57, étudiée sur les lots 7, 8 et 9 se montre plus hétérogène que la précédente. Le tracé des contours polygonaux convexes regroupant tous les individus du même lot facilite l'interprétation des relations. Ainsi il apparaît clairement que les lots 8 et 9, composés de femelles sont totalement séparés. Cette séparation provient-elle de la formation même des lots, ou d'un traitement antérieur ? Nous ne le savons pas, mais dans ce cas, ce traitement aurait eu un effet très net sur la morphologie des rongeurs en question.

Et c'est à ce propos que l'on peut apprécier un avantage énorme de l'analyse factorielle des correspondances sur les autres méthodes analogues : il est possible de confondre l'espace caractérisant les individus et celui des caractères. La projection simultanée des individus et des caractères sur le même plan est donc possible. Cela permet de mettre en évidence que tel lot est associé à tels caractères alors qu'un autre sera associé à d'autres caractères. On met en évidence les caractères qui ont pu être modifiés par le traitement dans le cas, bien entendu, du bien fondé de notre hypothèse initiale.

Résumé des observations.

La méthode utilisée a permis de mettre en évidence un certain nombre d'erreurs de mesures. Les lots SEC sont remarquablement homogènes, alors que les lots C 57 sont plus dispersés, indiquant une plus grande variabilité morphologique.

Les lots BALB et CDR sont relativement homogènes alors que le lot S x C est très nettement scindé en deux groupes.

Au sein des mêmes souches, les lots de même sexe sont séparés. Ce fait traduit une différence significative entre les lots considérés, qui peut provenir soit d'une différence de choix des individus dans la formation du lot, soit d'un traitement antérieur quelconque.

Conclusion.

L'analyse statistique rigoureuse de ce type de problème : discrimination entre lots, est relativement aisée puisqu'elle ne nécessite plus les fréquentes transformations des données brutes en vue de normaliser les distributions. Le travail du chercheur est ainsi considérablement allégé, la précision et la clarté de présentation des résultats notablement augmentée.

La méthode permet de traiter non seulement des données quantitatives mais aussi des observations qualitatives ce qui rend son emploi absolument universel, à condition que le nombre de caractères étudiés par sujet soit suffisamment élevé : dix paraît être un minimum. Il n'y a pas de limite supérieure autre que la capacité de l'ordinateur utilisé.

Manuscrit reçu au S.C.D. le 14 février 1972.

BIBLIOGRAPHIE

a) Concernant l'Analyse Factorielle des Correspondances :

L'essentiel des cours et publications s'y rapportant, peuvent être obtenus au Laboratoire de Statistique Mathématique, Tour 45, Faculté des Sciences, quai Saint-Bernard, Paris 5e.

Citons principalement :

Benzécri (J. P.) et Collaborateurs. - « L'analyse des données ». — Tome 1. La taxinomie, 460 p. — Tome 2. L'analyse des correspondances, 460 p. Parution : avril 1973, chez Dunod, Edit., 24-26, bd de l'Hôpital, Paris (5e).

Benzécri (J. P.), 1963-1964 – Cours de Linguistique Mathématique. 3e et 4e leçons. Multigr. Rennes.

Benzécri (J. P.), 1968 – Leçons sur l'Analyse Factorielle et la Reconnaissance des formes. Cours de 3e cycle. Multigr. Université de Paris.

Benzécri (J. P.), 1968 – 4e et 5e leçons : Analyse Factorielle et Analyse Factorielle des Correspondances. Cours de Statistique mathématique. Multigr. Université de Paris.

Benzécri (J. P.), 1968 – La Classification dans les Sciences de la Nature. Multigr.

Benzécri (J. P.), 1970 – Problème et méthodes de la Taxinomie. *Rev. Stat. Appl.*, vol. XVIII, n° 4, p. 73-98.

Benzécri (J. P.), 1971 – Sur les Algorithmes de Classification. *Rev. Stat. Appl.*, vol. XIX, n° 1, : 17-26.

Cordier (B.), 1965 – L'Analyse Factorielle des Correspondances. Thèse de 3e cycle.

b) Concernant l'application de l'Analyse Factorielle des Correspondances à la biométrie des rongeurs :

Bellier (L.), 1971 – Etude de neuf lots de Souris blanches par la méthode de l'Analyse Factorielle des Correspondances. Rapport multigr. 9 pages.

Bellier (L.), 1971 – Application de l'Analyse Factorielle des Correspondances à la biométrie des Rongeurs : séparation des *Cricetomys emini* et des *C. gambianus* de Côte d'Ivoire. Multigr. 37 pages.