MATHÉMATIQUES ET SCIENCES HUMAINES

B. LE ROUX
H. ROUANET

L'analyse multidimensionnelle des données structurées

Mathématiques et sciences humaines, tome 85 (1984), p. 5-18
<http://www.numdam.org/item?id=MSH_1984__85__5_0>

Math. Sci. hum. (22e année, n°85, *1984*, p.5-18)

L'ANALYSE MULTIDIMENSIONNELLE DES DONNEES STRUCTUREES[x]

B. LE ROUX[xx], H. ROUANET[xx]

RESUME

Le but de cet article est de montrer comment on peut , pour l'analyse des données d'observation , conjuguer les deux approches suivantes : d'une part les méthodes traditionnelles de l'analyse en composantes principales et ses variantes, d'autre part les méthodes d'analyse des données expérimentales comme l'analyse de la variance et, plus généralement , l'analyse des comparaisons . Nous suggérons alors, comme technique standard , de procéder à une double décomposition des inerties : selon chaque source de variation et selon chaque variable principale.

I - *INTRODUCTION*

Les données que nous nous proposons d'analyser seront des *données d'observation* provenant , par exemple, d'une enquête dans laquelle on pourra distinguer les *variables explicatives* des *variables à expliquer* . La distinction peut relever du plan d'enquête lui même (descripteurs des sujets comme l'âge,la catégorie socio-professionnelle ...) ou se faire au moment de l'analyse . Par analogie avec le langage de l'expérimentation, on pourrait dire "variables indépendantes" et "variables dépendantes" sinon "par construction" du moins "par intention" . Les données d'observation sont en général traitées par des procédures particulières comme l'*analyse en composantes principales* ou l'*analyse des correspondances* qui laissent de côté les concepts classiques de l'*analyse des données expérimentales* tels que par exemple celui d'*interaction* . Le but de cet article

---

[x] *L'essentiel de ce texte a fait l'objet d'une communication au congrès international de Psychométrie de Jouy-en-Josas(Juillet 1983)*

[xx] *Groupe Mathématique et Psychologie .*
Sciences Humaines - Sorbonne
Université René Descartes - 12, Rue Cujas, 75005 Paris

sera de montrer en quoi les méthodes issues de l'analyse des données expérimentales peuvent enrichir l'analyse des données d'observation .

Les données réelles obtenues à partir des procédures d'observation se réduisent rarement à un simple tableau à double entrée . Elles sont généralement constituées à partir de plusieurs facteurs , au sens où on parle de facteur en analyse de la variance . Les relations entre les facteurs engendrent une structure plus ou moins complexe mais formellement équivalente à celle d'un plan d'une véritable expérience , et qu'il nous parait essentiel de prendre en compte dans l'analyse des données . C'est pourquoi de telles données seront désormais appelées *données structurées* . Les relations entre les facteurs sont souvent très simples , beaucoup plus simples que celles rencontrées pour les données du laboratoire ; (elles se bornent souvent à celles de croisement et d'emboîtement); par contre , les *"plans" ne sont pas , en général , orthogonaux* . Il en résulte que le transfert des techniques élaborées dans un contexte expérimental sera assez facile , tandis que l'interprétation des résultats requerra beaucoup de soin .

Pour l'analyse des données structurées , un outil utile sera la notion formalisée de *comparaison* . Conceptuellement , la notion de comparaison étend celle de source de variation familière en analyse de variance et vise à répondre à toutes sortes de *questions spécifiques* se posant à propos des données , dans le cadre de la structure engendrée par le plan . Pour chaque question , on peut trouver une comparaison , à laquelle on associe une somme des carrés ou *inertie*. Cette méthode , qui constitue donc une extension de l'analyse de la variance classique sera appelée *analyse des comparaisons* à la suite de H.Rouanet et D.Lépine [5] .

Les inerties qui figurent dans la décomposition des sources de variation peuvent servir à des analyses descriptives aussi bien qu'inférentielles . Seules des analyses descriptives seront présentées ici : elles auront pour but de trouver des indices appropriés permettant d'évaluer la grandeur et l'importance des divers effets principaux des facteurs , de leur interaction etc... Ces indices descriptifs donneront lieu à des interprétations comparables à celles des contributions en analyse des correspondances . L'utilisation à la fois de l'analyse des comparaisons et de l'analyse en composantes principales nous conduit naturellement à une *double décomposition des inerties* selon chaque source de variation et chaque variable principale .

II - *ANALYSE DES COMPARAISONS*

Nous donnerons maintenant les définitions des notions fondamentales de l'analyse des comparaisons : celles de contraste , de comparaison et d'inertie .[5]

II.1 - *Contraste et comparaison*

Soit J un ensemble fini .

On appelle *contraste sur J* toute mesure sur J dont la somme des coefficients (ou masse totale) est nulle .

On appelle *comparaison sur J* tout sous-espace vectoriel de contrastes sur J .

En particulier , nous parlerons de la comparaison à 1 degré de liberté (1 d.l.) engendrée par un contraste ; de la comparaison à p d.l. ($p \geq 1$) engendrée par une base de p contrastes etc... L'espace de tous les contrastes sur J sera appelé la *comparaison globale sur J* .

Etant donnée une mesure fondamentale strictement positive sur J , l'espace vectoriel des mesures sur J est muni de la structure euclidienne selon laquelle la norme du contraste c_J est $(\Sigma c_j^2/n_j)^{1/2}$.

II.2 - *Nuage euclidien*

Nous allons maintenant appliquer les notions précédentes à un nuage pondéré de points d'un *espace affine euclidien* , noté (M^J, n_J) , de point moyen (ou barycentre) $G = \Sigma n_j M^j / \Sigma n_j$.

Rappelons d'abord la notion classique d' *inertie d'un nuage* :

$$\Sigma n_j (GM^j)^2$$

où GM^j désigne la distance euclidienne entre les points G et M^j .

II.3 - *Effet d'un contraste sur un nuage et inertie associée* [4]

Nous définirons *l'effet d'un contraste sur un nuage* M^J par le vecteur $\overrightarrow{\Sigma c_j M^j}$

Nous définirons *l'inertie (ou somme des carrés) associée* par :

$$\frac{\| \overrightarrow{\Sigma c_j M^j} \|^2}{\Sigma c_j^2 / n_j}$$

où le numérateur est le carré de la norme euclidienne du vecteur effet et le dénominateur le carré de la norme du contraste . Cette inertie est celle associée à tout contraste proportionel , donc nous la prendrons pour définir l'inertie de la comparaison (à 1 d.l.) engendrée par ce contraste .

Nous définirons enfin l'inertie associée à une comparaison à p

degrés de liberté , avec $p \geq 1$. Cette comparaison sera représentée par une base de p contrastes . L'inertie associée pourra être définie comme la somme des inerties associées à p contrastes orthogonaux . Bien entendu , cette inertie ne dépend pas du choix de la base .

II.4 - *Croisement de deux facteurs*

Nous envisageons maintenant le cas particulier du croisement de deux facteurs A et B , noté AxB . Dans ce cas , on peut définir des types particuliers de comparaisons à savoir les comparaisons intra et d'interaction ainsi que des dérivations vers A et vers B permettant de construire à partir du nuage défini sur AxB des nuages dérivés sur A et sur B [4] .

Contrastes et comparaisons intra

Un contraste sur AxB est un *contraste intra-A* si pour chaque a de A c'est un contraste sur B :

$$\forall\ a \in A \quad \sum_B c_{ab} = 0$$

On définira de même un *contraste intra-B* .

Une *comparaison intra-A* sera engendrée par une famille de contrastes intra-A .

Contrastes et comparaisons d'interaction

Un contraste sur AxB est un *contraste d'interaction* entre A et B s'il est à la fois contraste intra-A et contraste intra-B .

$$\forall\ a \in A \quad \sum_B c_{ab} = 0$$

$$\forall\ b \in B \quad \sum_A c_{ab} = 0$$

En particulier le produit terme à terme d'un contraste sur A et d'un contraste sur B est un contraste d'interaction . On démontre facilement que , étant donné une base de contrastes sur A et une base de contrastes sur B , la famille des contrastes-produits est une base de la comparaison globale d'interaction qui sera notée A.B .

Dérivation d'un nuage

On étudiera le cas de la dérivation vers A .

Etant donné une transition τ_B^A de A vers B , on définit le *nuage dérivé sur A* (noté M^A) par :

$$\forall\, a \in A \qquad M^a = \sum_B \tau_b^a \, M^{ab}$$

Le point M^a est le barycentre des points $(M^{ab})_{b \in B}$ affectés des poids $(\tau_b^a)_{b \in B}$. L'effet global du facteur A dépendra donc de la transition τ_B^A . Le nuage M^A n'est pas , à priori , muni d'une pondération . Dans [4] on a précisé pour divers types de transition τ_B^A les pondérations compatibles . En particulier , pour la *dérivation pondérée par les effectifs* c'est à dire pour $\tau_B^A = (n_{ab}/n_a)_{a \in A, b \in B}$, on pourra munir le nuage M^A de la pondération n_A et ainsi calculer l'inertie associée à la source de variation A : elle vaut $\Sigma\, n_a\, (GM^a)^2$.

Décomposition de l'inertie selon les sources de variation

En *analyse de la variance* , il est d'usage de décomposer la source de variation AxB en les trois sources A , B et l'interaction A.B . Nous ferons de même en *analyse des comparaisons* , quand le plan est équilibré ou plus généralement orthogonal , les comparaisons associées à ces trois sources de variation sont orthogonales et les inerties sont additives .

En *analyse des données structurées* , le plan n'est pas en général orthogonal , il n'y aura donc pas décomposition additive des inerties , mais ceci n'empêche pas de procéder à cette décomposition et surtout d'interpréter les effets d'interaction vis à vis des effets principaux .

III - *LA DOUBLE DECOMPOSITION DES INERTIES*

Les données d'observation sont communément analysées par des méthodes multivariées familières comme l'analyse en composantes principales ou l'analyse des correspondances ; il sera souvent intéressant , pour des données structurées , de combiner les deux approches . Pour ce faire , nous suggérons , comme technique standard , de procéder systématiquement à une double décomposition des inerties selon chaque source de variation d'une part et chaque variable principale d'autre part .

IV - *EXEMPLE*

Nous appliquerons la méthode d'analyse des données structurées à des *données psychométriques* empruntées à Rouanet , Lépine et Pelnard [6] . Ces données proviennent d'une enquête effectuée sur 334 élèves répartis en 4 groupes ayant à peu près le même effectif . Ces groupes sont obtenus en croisant deux facteurs A et B ayant chacun deux modalités .

A : Méthode d'enseignement avec a moderne a' traditionnel

B : Milieu avec b favorisé b' défavorisé

		Favorisé b	Défavorisé b'
Moderne	a	86	86
Traditionnel	a'	80	82

Les variables dépendantes étaient 9 tests de développement intellectuel et de connaissance à savoir QI verbal , QI non verbal , 2 tests de combinatoire, 2 tests de probabilité , 2 tests de logique des propositions et un test de maths .

Nous prendrons ici le point de vue multidimensionnel et nous nous bornerons à des analyses descriptives(*). Après avoir procédé à une réduction variable par variable , nous ferons correspondre à chaque élève un profil (9 valeurs numériques) qui sera représenté par un point dans un espace affine muni de la métrique élémentaire .

IV.1 - *Analyse multidimensionnelle descriptive des comparaisons*

Pour calculer les vecteurs-effets relatifs aux sources de variation A , B et A.B , on procèdera à deux dérivations successives: la première consiste à passer du nuage initial au nuage dérivé M^{AB} qui comporte les quatre points (M^{ab},$M^{ab'}$,$M^{a'b}$,$M^{a'b'}$) ; la deuxième consiste à obtenir les

(*) : Le point de vue inférentiel étudiant variable après variable est exposé dans [6] ; l'analyse inférentielle multivariée pourrait être faite à partie de [3] .

points moyens (M^a, $M^{a'}$) et (M^b, $M^{b'}$) définis comme barycentres des points du nuage M^{AB} avec des pondérations appropriées : nous prendrons ici la dérivation pondérée par les effectifs (cf fig 1) .

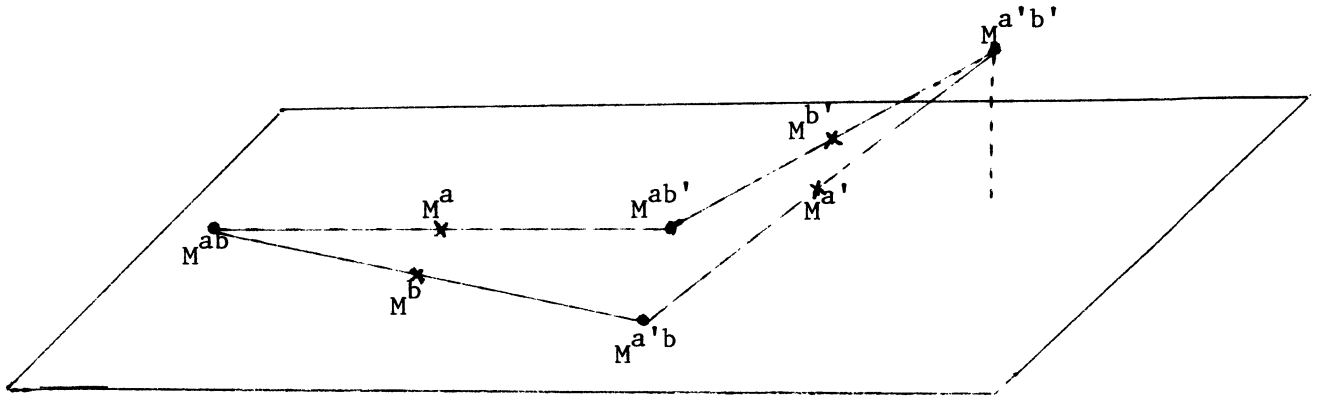

Figure 1 : Représentation du nuage M^{AB} et des points moyens $(M^a, M^{a'}), (M^b, M^{b'})$

Le vecteur effet associé à A est $\overrightarrow{M^a - M^{a'}}$

à B est $\overrightarrow{M^b - M^{b'}}$

à l'interaction A.B est $\overrightarrow{M^{ab} - M^{a'b} - M^{ab'} + M^{a'b'}}$

Comme indice descriptif permettant d'évaluer la *grandeur des effets*, on pourra prendre la norme euclidienne des vecteurs effets . On obtient les résultats suivants :

source de variation	norme du vecteur effet
A	1.1365
B	1.1463
A.B	0.5251

Une autre façon d'apprécier la grandeur de l'effet serait de construire , à partir du nuage M^{AB} , au moyen d'une dérivation appropriée un *nuage additif* - i.e. sans interaction - et d'évaluer la grandeur des effets d'interaction par une norme moyenne des écarts vectoriels, en d'autres termes des *résidus* (vectoriels) par rapport au nuage additif. Une telle norme moyenne sera environ (exactement s'il y a équipondération) le quart de la

norme de l'effet d'interaction défini ci-dessus . Quant à *l'importance de l'effet d'interaction* , elle pourra être appréciée descriptivement par un rapport de normes . Par exemple , si on fait le rapport de la norme des écarts au modèle additif à la norme de l'un ou l'autre des effets principaux , on trouvera ici un rapport de l'ordre de 1/8 .

Puis on calcule , comme en analyse de la variance , *les inerties associées à chaque source de variation* . On obtient les résultats suivants :

source de variation	inertie	d.l.
AxB	222.13	3
A	108.04	1
B	109.47	1
A.B	5.78	1
S(AxB)	2783.88	330

De plus , on constate , en calculant les angles entre les vecteurs effets associés aux trois sources de variation que le vecteur effet associé à A.B n'est pas dans le plan des vecteurs effets associés à A et à B . Ce dernier point sera repris au cours de l'étude conjointe des résultats de l'analyse des comparaisons et de l'analyse en composantes principales .

IV.2 - Analyse en composantes principales

Effectuons maintenant l'analyse en composantes principales du nuage des 334 points . Nous obtenons , pour les variances des quatre premières variables principales , les résultats suivants :

	1	2	3	4
valeur propre	4.606	0.937	0.770	0.626
% de variance	51.14	10.42	8.55	6.96

Nous effectuerons des représentations graphiques dans le plan des deux premiers axes principaux (56 % de la variance totale du nuage), à savoir les projections des points des sous-nuages correspondants aux 4 groupes (fig. 2) , puis les projections des 4 points M^{AB} ainsi que celles de (M^{a},$M^{a'}$) et de (M^{b},$M^{b'}$) (fig. 3) .

Le *premier axe* oppose pédagogie moderne (a) et milieu favorisé (b) d'une part à pédagogie traditionnelle (a') et milieu défavorisé (b') d'autre part .

Le *deuxième axe* montre une opposition peu marquée entre pédagogie moderne (a) et milieu défavorisé (b') d'une part et pédagogie traditionnelle (a) et milieu favorisé (b') , ce qui revient à dire que le deuxième axe prend en compte une partie de l'interaction .

IV.3 Double décomposition des inerties

Si nous procédons maintenant à la *double décomposition des inerties* selon chaque *source de variation* et chaque *variable principale* , on obtient :

source de variation	inertie totale	variables principales 1	2	3	4
AxB	222.13	192.55	3.57	7.46	1.66
A	108.04	86.73	0.75	6.84	0.57
B	109.47	104.85	0.83	0.38	0.30
A.B	5.78	2.14	1.97	0.25	0.79

Les inerties apparaissant dans ce tableau se calculent facilement à l'aide du programme *VAR UNI G* réalisé en 1978 par M.O. Lebeaux [2] .

A partir de ce tableau nous pouvons étudier les variables une par une . Pour chaque axe principal , nous calculerons

sa *contribution à l'inertie* associée à une *source de variation* : cette contribution est aussi le carré du cosinus de l'angle que fait le vecteur effet avec l'axe principal ; c'est la généralisation , pour l'analyse des données structurées, de la *contribution de l'axe au point* , indice classique en analyse des correspondances [1] .

Tableau des contributions des axes
à l'inertie associée aux sources de variation

source de variation	axe 1	axe 2	axe 3	axe 4
AxB	86.68	1.61	3.36	0.74
A	80.27	0.69	6.33	0.53
B	95.78	0.76	0.35	0.27
A.B	37.01	34.17	4.32	13.66

A partir du tableau des contributions , on voit immédiatement que le 1er axe contribue presque complètement à l'inertie associée à A et à B ; par contre , il faut au moins les deux premiers axes, auquel il serait bon d'adjoindre le 4ème, pour prendre en compte une bonne part de l'inertie d'interaction. On voit ainsi que si, dans l'analyse en composantes principales faite ici, on se contentait de prendre en compte les deux premiers axes principaux, même s'ils contribuent à la majeure partie de l'inertie totale, on négligerait une bonne partie des effets d'interaction.

Dans cet article , nous avons présenté un exemple portant sur des variables numériques , la structure des données étant du type $S\langle AxB\rangle \longrightarrow R^n$, *mais l'analyse des données structurées peut également traiter un tableau de contingence ternaire* [8] *ou même multiple, de structure, par exemple,* $(AxB)x(CxD) \longrightarrow N$ *: on fera alors jouer un rôle dissymétrique aux variables, en prenant comme variables explicatives celles ,par exemple , liées au croisement AxB, et on étudiera alors l'interaction entre A et B relativement aux variables à expliquer ,ici C et D* [7] .

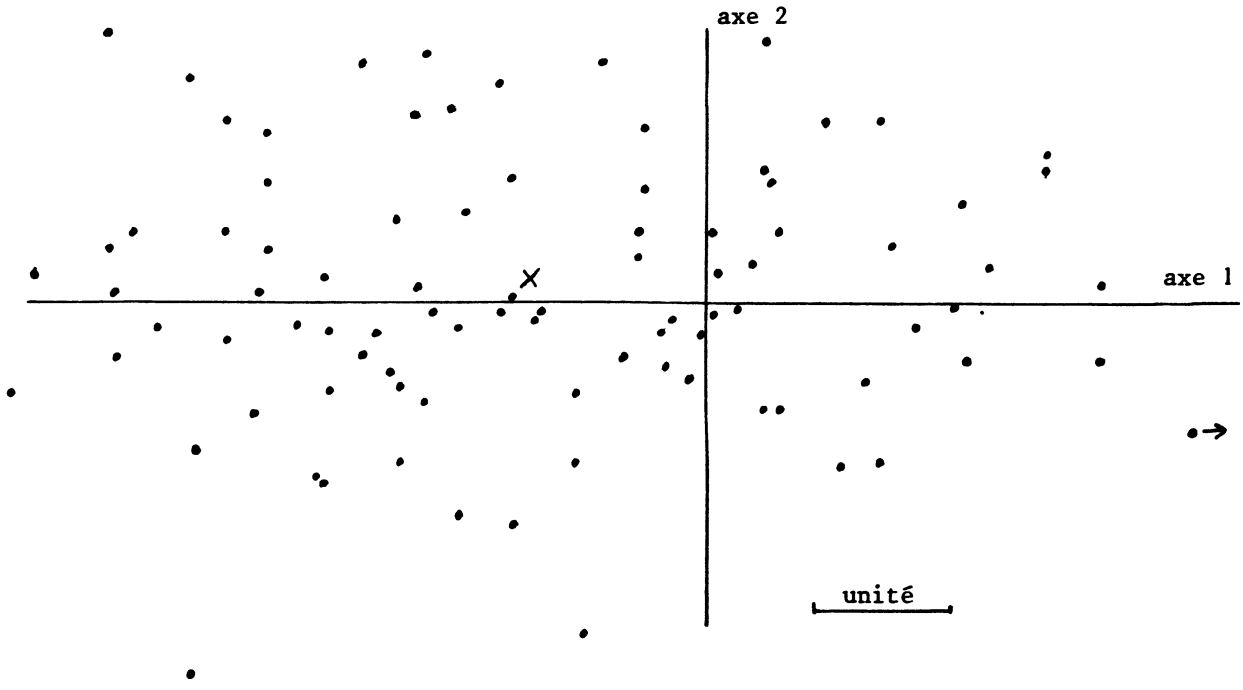

Fig. 2.1 : Plan des axes 1 et 2; sous-nuage correspondant au groupe (a,b)
a : pédagogie moderne ; b : milieu favorisé

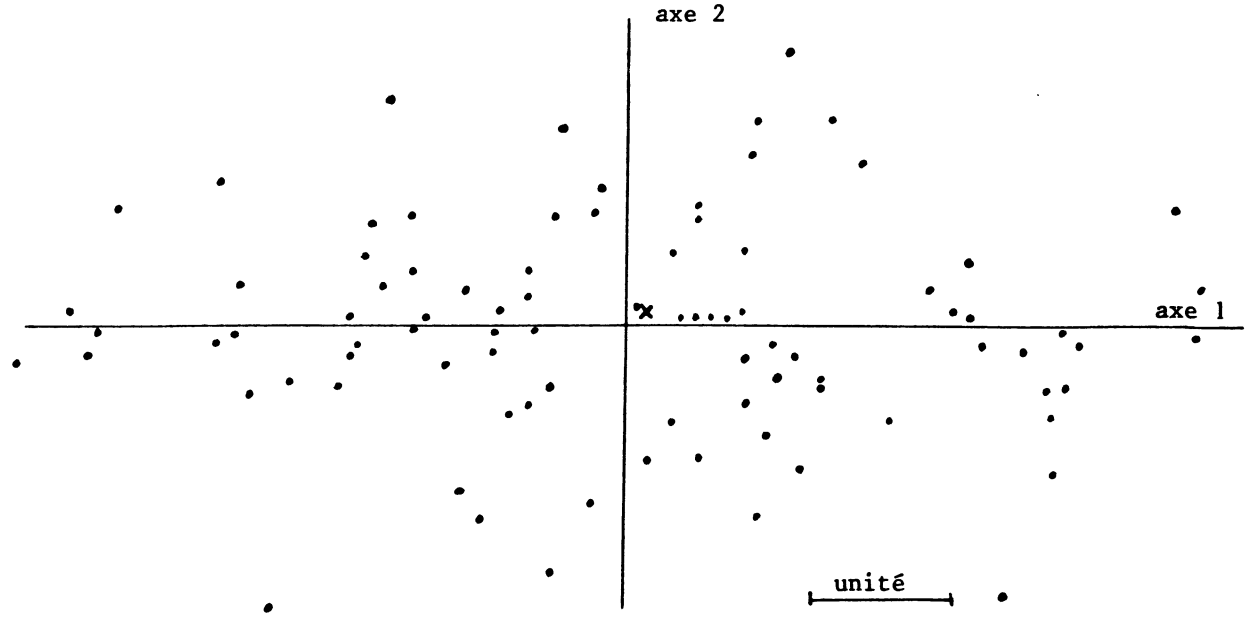

Fig. 2.2 : Plan des axes 1 et 2; sous-nuage correspondant au groupe (a,b')
a : pédagogie moderne ; b' : milieu défavorisé

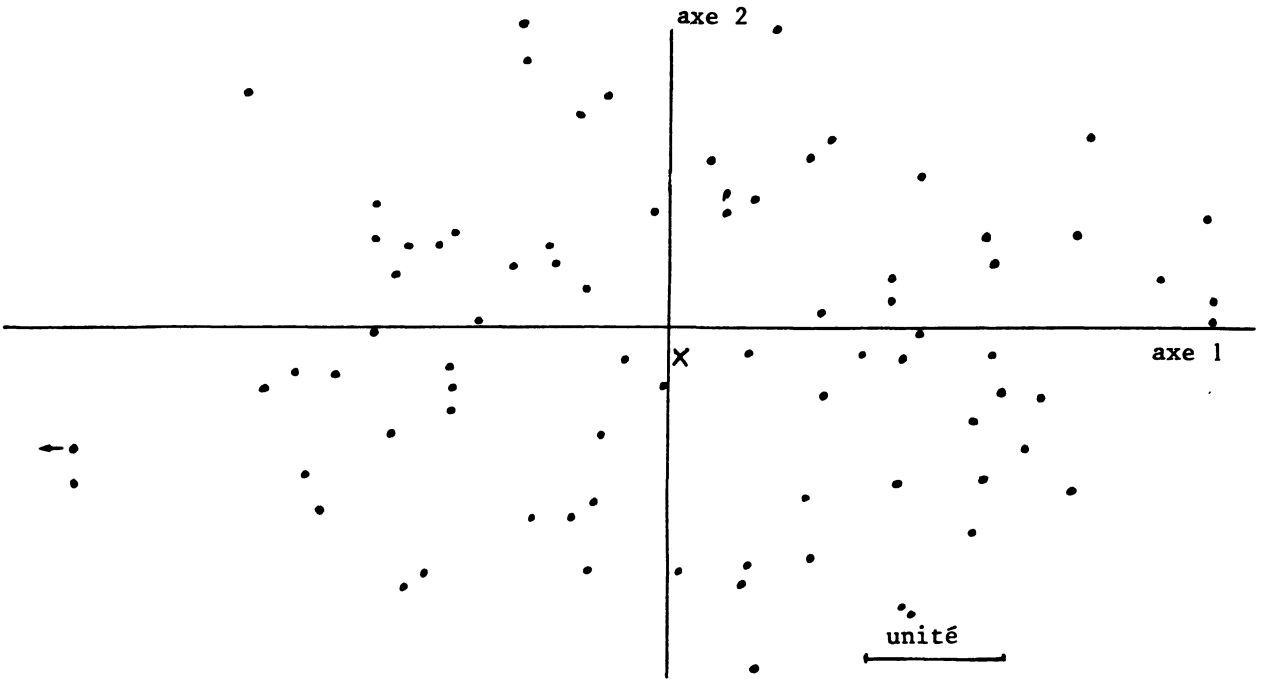

Fig. 2.3 : Plan des axes 1 et 2; sous-nuage correspondantau groupe (a',b)
a' : pédagogie traditionnelle ; b : milieu favorisé

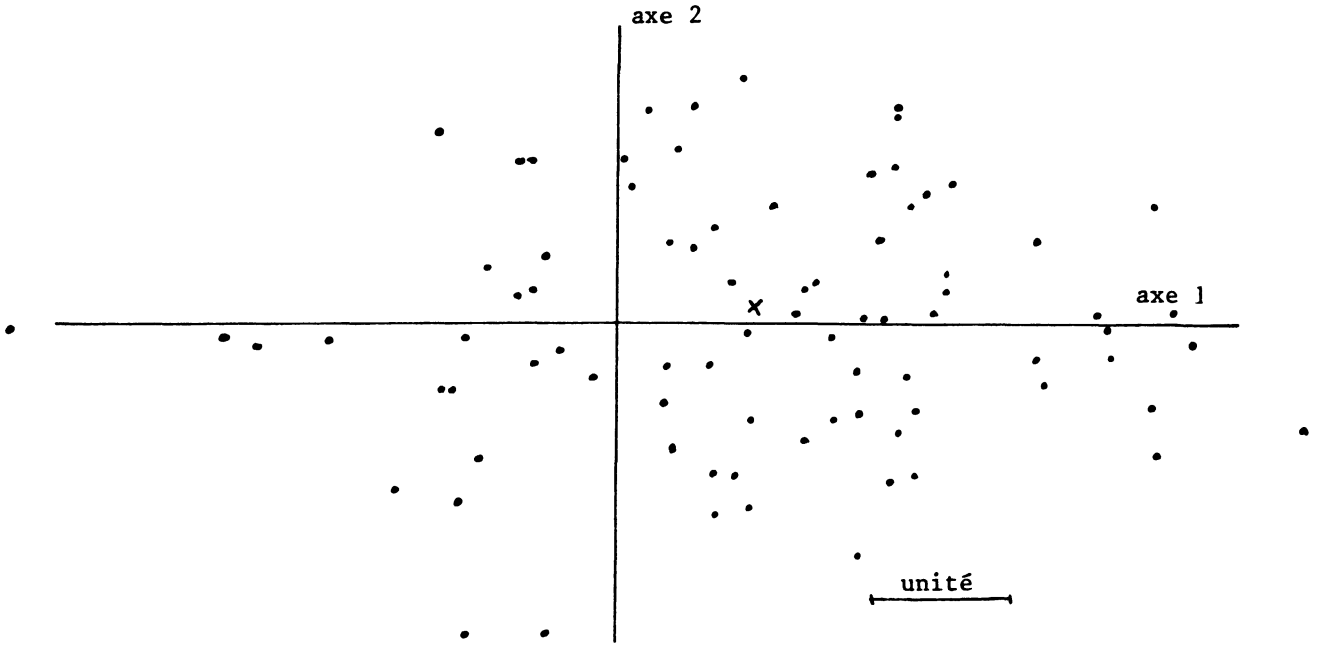

Fig. 2.4 : plan des axes 1 et 2; sous-nuage correspondant au groupe (a',b')
a' : pédagogie traditionnelle ; b' : milieu défavorisé

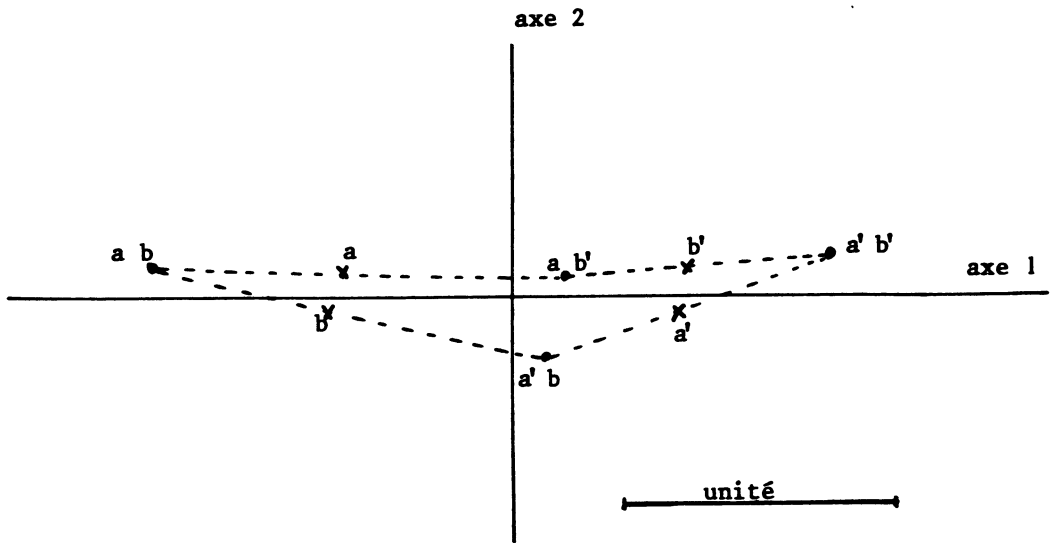

Fig. 3 : plan des axes 1 et 2 ; projections des 4 points du nuage M^{AB} et des points moyens $(M^{a},M^{a'})$ et $(M^{b},M^{b'})$ ..
a : pédagogie moderne , a' : pédagogie traditionnelle
b : milieu favorisé , b' : milieu défavorisé

BIBLIOGRAPHIE

[1] BENZECRI J.-P. & F.,*Pratique de l'analyse des données,tome 1,Analyse des correspondances,*Paris,Dunod,1980 .

[2] LEBEAUX M.-O. & ROUANET H.,*Notice d'utilisation du programme VAR UNI G,* texte multigraphié,1978,Groupe mathématique et psychologie .

[3] LECOUTRE B. & ROUANET H.,*Deux structures statistiques fondamentales en analyse de la variance univariée et multivariée,* Math.Sci.hum.,75,1981,p.71-82.

[4] LE ROUX B. & ROUANET H.,*L'analyse statistique des protocoles multidimensionnels : analyse des comparaisons,* Pub.Inst.Stat. Univ.,XXVIII,fasc.1,2,1983,47-70 .

[5] ROUANET H. & LEPINE D.,*Structures linéaires et analyse des comparaisons,* Math.sci.hum.,56,1976,p.5-46 .

[6] ROUANET H.,LEPINE D.,PELNARD-CONSIDERE J.,*Bayes-fiducial procedures as practical substitutes for misplaced significance testing: application to educational data.*Communication à l'"international Symposium for educational testing",Montreux,1975,in D.N.M. de Gruijter and L.J.Th. Van der Kamp (Eds) :Advances in Psychological and educational Measurement,J.Wiley and Sons,p.35-50.

[7] ROUANET H. Compte-rendu de fin d'étude (ATP 4214). Structuration des données d'observation et planification des analyses, 1982

[8] WALFARD D. Mémoire de DEA, Université René Descartes, 1979