MATHÉMATIQUES ET SCIENCES HUMAINES

C. DENIAU
G. OPPENHEIM

Deux méthodes linéaires en statistique multidimensionnelle (2). Analyse des tableaux de correspondances

Mathématiques et sciences humaines, tome 45 (1974), p. 5-28
<http://www.numdam.org/item?id=MSH_1974__45__5_0>

Math. Sci. hum. (12e année, n° 45, 1974, p. 5-28)

DEUX MÉTHODES LINÉAIRES EN STATISTIQUE MULTIDIMENSIONNELLE (2)

ANALYSE DES TABLEAUX DE CORRESPONDANCES

par

C. DENIAU et G. OPPENHEIM

RÉSUMÉ

Ce texte constitue la suite de l'article «Deux méthodes linéaires en statistique multidimensionnelle» paru dans le n° 44 de cette revue. Nous nous intéressons ici aux tableaux d'effectifs. La théorie du paragraphe 1.2 *est appliquée pour obtenir les résultats : détermination des composantes et axes principaux, construction des graphiques, indices, analyses conjointes des deux nuages associés au tableau des données.*

On insiste sur quelques difficultés courantes de l'interprétation des résultats.

Plusieurs problèmes dont certains méthodologiques sont effleurés : codages, rotations des axes, symétrie de deux ensembles, etc.

SUMMARY

This article is a follow-up to the article entitled, « Two linear methods in multi-dimensional statistics », which appeared in No. 44 *of this journal. Here we are interested in contingency tables. The theory of paragraph* 1.2 *is applied in order to obtain the following results : determination of components and principal axes, construction of graphs, indices, joint analysis of two clusters associated with tables of data.*

One insists on some current difficulties in interpreting results.

There are many problems where certain methodologies are merely touched upon : coding, rotation of axes, symmetry of two sets, etc.

1. INTRODUCTION ET EXEMPLE

1.1. Introduction

Ce texte constitue la suite de l'article « Deux méthodes d'analyse factorielle [1] » paru dans cette revue (n° 44). Les notions utiles et le problème à résoudre sont ceux des § 0 et 2 du précédent article. Dans ce qui suit nous particularisons les définitions au cas des tableaux de correspondances. Tous les résultats du § 2 sont applicables ici ; dans la plupart des cas, il sera suffisant de spécifier quelques définitions et de laisser au lecteur le soin de déduire les différentes propriétés dont aucune démonstration ne présente de difficultés importantes. Nous suivrons le plus souvent ce chemin, nous attachant à mettre en évidence les écueils.

1. Les références à des paragraphes de cet article sont précédées de I.

D'après J.-M. Faverge [15] les principes de cette méthode d'analyse de tableaux d'effectifs datent de 1950, Guttman [16]. D'autre part en 1952 C. Hayashi [17] et [18] a développé l'étude des tableaux de correspondances et introduit certaines représentations graphiques.

Un développement systématique a été réalisé en France par J.-P. Benzécri [11] permettant d'épurer les méthodes anciennes des hypothèses inoffensives ou inutiles en faisant mieux apparaître le cadre linéaire, et non probabiliste de la méthode. De plus la réalisation et la diffusion de programmes d'ordinateur a rendu possible l'utilisation de la méthode dans un grand nombre de disciplines.

1.2. Exemple d'analyse des correspondances

La courte illustration qui va suivre est tirée de l'*Etude de la composition par âges de* 141 *villes touristiques du littoral français* [13], auquel on se rapportera pour le détail des tableaux de données, des résultats, des interprétations, des discussions méthodologiques et du choix des villes ; ces éléments trop nombreux et trop volumineux ne peuvent trouver place dans cet exemple.

a) *Les données* : Chacune de ces 141 villes touristiques littorales (V.T.L.) est décrite par sa composition par âge. Pour chaque sexe le recensement de 1968 nous donne des classes dont l'amplitude est 5 ans (0 à 5 ans, 5 ans à 10 ans, etc.), et une classe pour les plus de 75 ans. Par abus de langage on parlera de 32 classes d'âges (16 classes d'âges par sexe).

i) Ainsi à chacune des 141 V.T.L. est associée une suite de 32 nombres entiers positifs ou nuls. Chaque nombre est l'effectif de la population masculine ou féminine d'une classe d'âge de cette ville. Le profil (ou description) d'une ville est la distribution des fréquences des 32 classes d'âges de cette ville.

ii) A chacune des 32 classes d'âges est associée une suite de 141 nombres entiers positifs ou nuls ; pour une classe chaque nombre est l'effectif de cette classe d'âge dans une des 141 villes. Le profil (ou description) d'une classe d'âge est la distribution des fréquences des 141 villes pour cette classe.

b) *L'analyse* : Les tableaux et graphiques sont les suivants :

Tableau 1 : Code des V.T.L. et des classes d'âges.
Tableau 2 : Effectifs par âge et sexe de la population des 141 V.T.L.
Tableau 3 : Valeurs propres et qualité de l'ajustement.
Tableau 4 : Résultats concernant les classes d'âges pour les 4 premiers axes.
Figure 1 : Les classes d'âges et les V.T.L. dans le plan (1, 2).
Figure 2 : Cartographie des projections des V.T.L. sur l'axe 2.
Figure 3 : Les classes d'âges dans le plan (3, 4).

Tableau 1. — *Code de quelques villes touristiques littorales.*

A 01	BRAY-DUNES	H 04	PLESTIN	X .2	MARTIGUES
A .1	MALO-LES-BAINS	H 05	SAINT-QUAY-PORTRIEUX	Y 01	BANDOL
B .1	BERCK	I .1	CONCARNEAU	Y .1	FREJUS VILLE
B 01	EQUIHEN	I 11	ROSCOFF	Y 04	LA LONDE
B .2	LE PORTEL	J 01	AURAY	Y 07	SAINTE-MAXIME
B 03	WIMEREUX	J 02	LARMOR-PLAGE	Y .3	SAINT-RAPHAEL
C 01	CAYEUX	K .1	LA BAULE	Y 09	SAINT-TROPEZ
C 02	MERS-LES-BAINS	K 05	PORNICHET	Y 10	SANARY
D .1	DIEPPE	K 07	SAINT-BREVIN	Y .4	LA SEYNE
D .2	FECAMP	L 01	ILE D'YEU	Z .1	ANTIBES
D 02	LE TREPORT	L .1	SABLES-D'OLONNE	Z 01	BEAULIEU
E 01	CABOURG	M .1	ROYAN	Z .2	BEAUSOLEIL
E 03	DIVES-SUR-MER	N .1	ARCACHON	Z .3	CAGNES-SUR-MER
E 04	HONFLEUR	N 05	SOULAC-SUR-MER	Z .4	CANNES
E 10	TROUVILLE	N .2	LA TESTE	Z .5	LE CANNET
F 01	DEAUVILLE	O 01	BISCAROSSE	Z 02	CAP-D'AIL
F .1	GRANVILLE	O 02	CAPBRETON	Z .6	MENTON
F 02	SAINT-CAST	O 03	MIMIZAN	Z 04	ROQUEBRUNE
G 01	CANCALE	P .2	BIARRITZ	Z 05	SAINT-JEAN-CAP-FERRAT
G 02	DINARD	T 02	BANYULS	Z .7	SAINT-LAURENT-DU-VAR
G .1	SAINT-MALO	T 04	COLLIOURE	Z .8	VALLAURIS
H 03	PLENEUF	W 01	GRAU-DU-ROI	Z 06	VILLEFRANCHE

Tableau 2. — *Effectifs par âge et sexe de l'ensemble de la population des 141 villes touristiques littorales*

Hommes	Age		Femmes
24 649	75 et plus		48 220
20 853	70 ≤	< 75	32 560
29 730	65 ≤	< 70	39 256
33 745	60 ≤	< 65	41 731
35 778	55 ≤	< 60	41 837
26 332	50 ≤	< 55	29 727
38 849	45 ≤	< 50	42 207
40 841	40 ≤	< 45	42 470
41 570	35 ≤	< 40	42 037
37 434	30 ≤	< 35	37 922
34 268	25 ≤	< 30	33 997
42 231	20 ≤	< 25	43 650
50 733	15 ≤	< 20	49 680
50 821	10 ≤	< 15	48 827
51 746	5 ≤	< 10	50 056
42 598	0 ≤	< 5	40 992

Tableau 3

Axes	Valeurs propres associées	Qualité de l'ajustement en pourcentage	Qualité cumulée en pourcentage
1	0,169	59,90	59,90
2	0,039	13,90	73,80
3	0,017	6,04	79,84
4	0,014	4,98	84,82
5	0,006	2,12	86,94
6	0,0055	1,98	89,01
7	0,0038	1,35	90,36
8	0,0025	0,88	91,24
9	0,0022	0,77	92,01
10	0,0021	0,75	92,76

CODE DES CLASSES D'AGES

Une lettre M (masculin) ou F (féminin) précédant un multiple de 5. Ainsi :
M 25 représente la classe des hommes tels que 25 ≤ âge < 30
F 50 représente la classe des femmes tels que 50 ≤ âge < 55
Pour M 75 et F 75 il s'agit d'individus de 75 ans et plus.

Tableau 4. — *Résultats concernant les classes d'âge pour les quatre premiers axes*

	1			2			3			4		
	a	b	c	a	b	c	a	b	c	a	b	c
		%			%			%			%	
M 00	— 0,1661	5,48	0,786917	— 0,0045	0,02	0,000577	— 0,0514	5,20	0,075219	0,0408	3,98	0,047444
M 05	— 0,1466	5,18	0,756058	— 0,0155	0,25	0,008433	— 0,0608	8,84	0,130069	— 0,0247	1,77	0,021398
M 10	— 0,1192	3,36	0,645011	— 0,0647	4,27	0,190033	— 0,0152	0,54	0,010438	— 0,0272	2,10	0,033464
M 15	— 0,0798	1,51	0,302274	— 0,1067	11,60	0,539986	0,0153	0,55	0,011142	— 0,0004	0,00	0,000008
M 20	— 0,0582	0,67	0,236813	— 0,0306	0,79	0,065270	0,0437	3,73	0,133424	0,0603	8,61	0,253932
M 25	— 0,0645	0,67	0,129537	0,1296	11,56	0,522343	0,0073	0,09	0,001672	0,0852	13,96	0,225655
M 30	— 0,0812	1,15	0,262200	0,1183	10,52	0,556617	— 0,0300	1,56	0,035859	0,0086	0,15	0,002917
M 35	— 0,0683	0,90	0,256725	0,0963	7,74	0,510590	— 0,0149	0,42	0,012164	— 0,0336	2,64	0,062192
M 40	— 0,0371	0,26	0,095924	0,0577	2,73	0,231800	0,0164	0,51	0,018651	— 0,0672	10,36	0,314456
M 45	— 0,0068	0,01	0,005271	0,0071	0,04	0,005803	0,0666	7,97	0,508904	— 0,0248	1,34	0,070446
M 50	0,0362	0,16	0,090000	0,0276	0,40	0,052307	0,0671	5,48	0,308735	— 0,0161	0,38	0,017861
M 55	0,0839	1,17	0,431385	0,0078	0,04	0,003722	0,0450	3,35	0,124128	— 0,0236	1,12	0,034190
M 60	0,1303	2,67	0,663196	0,0319	0,69	0,039877	0,0091	0,13	0,003271	— 0,0196	0,72	0,014937
M 65	0,1896	4,97	0,770716	0,0339	0,69	0,024643	— 0,0356	1,74	0,027214	— 0,0201	0,67	0,008662
M 70	0,2493	6,03	0,811876	0,0249	0,26	0,008092	— 0,0554	2,96	0,040155	— 0,0051	0,03	0,000345
M 75	0,2998	10,32	0,851356	0,0161	0,13	0,002465	— 0,0704	5,64	0,046917	0,0165	0,38	0,002582
F 00	— 0,1682	5,40	0,782566	— 0,0146	0,17	0,005876	— 0,0553	5,79	0,084524	0,0326	2,44	0,029336
F 05	— 0,1396	4,54	0,756418	— 0,0295	0,87	0,033776	— 0,0454	4,77	0,080018	— 0,0245	1,69	0,023347
F 10	— 0,1222	3,40	0,623911	— 0,0671	4,41	0,187741	— 0,0208	0,97	0,018032	— 0,0418	4,78	0,072833
F 15	— 0,0877	1,78	0,277371	— 0,1267	16,01	0,578580	0,0312	2,23	0,034995	0,0093	0,24	0,003094
F 20	— 0,0589	0,71	0,186213	— 0,0220	0,42	0,025966	0,0519	5,44	0,144505	0,0943	21,78	0,476623
F 25	— 0,0584	0,54	0,142991	0,1141	8,88	0,545413	0,0012	0,00	0,000061	0,0604	6,96	0,152901
F 30	— 0,0660	0,77	0,254977	0,0960	7,02	0,539648	— 0,0111	0,21	0,007181	— 0,0068	0,10	0,002698
F 35	— 0,0382	0,29	0,150044	0,0580	2,84	0,344607	0,0089	0,15	0,008153	— 0,0370	3,23	0,140402
F 40	0,0089	0,02	0,009868	0,0186	0,29	0,042626	0,0491	4,73	0,298111	— 0,0402	3,85	0,199696
F 45	0,0354	0,25	0,133871	— 0,0125	0,13	0,016851	0,0683	9,11	0,499316	0,0002	0,00	0,000005
F 50	0,0842	0,98	0,429186	0,0052	0,02	0,001633	0,0663	6,04	0,266423	— 0,0130	0,28	0,010166
F 55	0,1242	3,01	0,705548	— 0,0047	0,02	0,001023	0,0342	2,26	0,053412	— 0,0053	0,07	0,001277
F 60	0,1562	4,74	0,830703	— 0,0053	0,02	0,000969	0,0070	0,09	0,001665	— 0,0097	0,22	0,003212
F 65	0,2118	8,21	0,875568	— 0,0190	0,29	0,007067	— 0,0373	2,53	0,027200	— 0,0068	0,10	0,000895
F 70	0,2326	8,20	0,836826	— 0,0452	1,34	0,031665	— 0,0482	3,50	0,035954	0,0102	0,19	0,001601
F 75	0,2373	12,65	0,797928	— 0,0757	5,55	0,081206	— 0,0394	3,46	0,022003	0,0465	5,85	0,030635

a : coordonnées des projections de la classe d'âge sur l'un des axes (factor-scores).

b : pourcentage de la dispersion totale de l'axe prise en compte par la classe d'âge.

c : qualité de la représentation de la classe d'âge par l'axe.

c) Quelques résultats de l'analyse empruntés à [13] (pour un exposé complet s'y reporter).

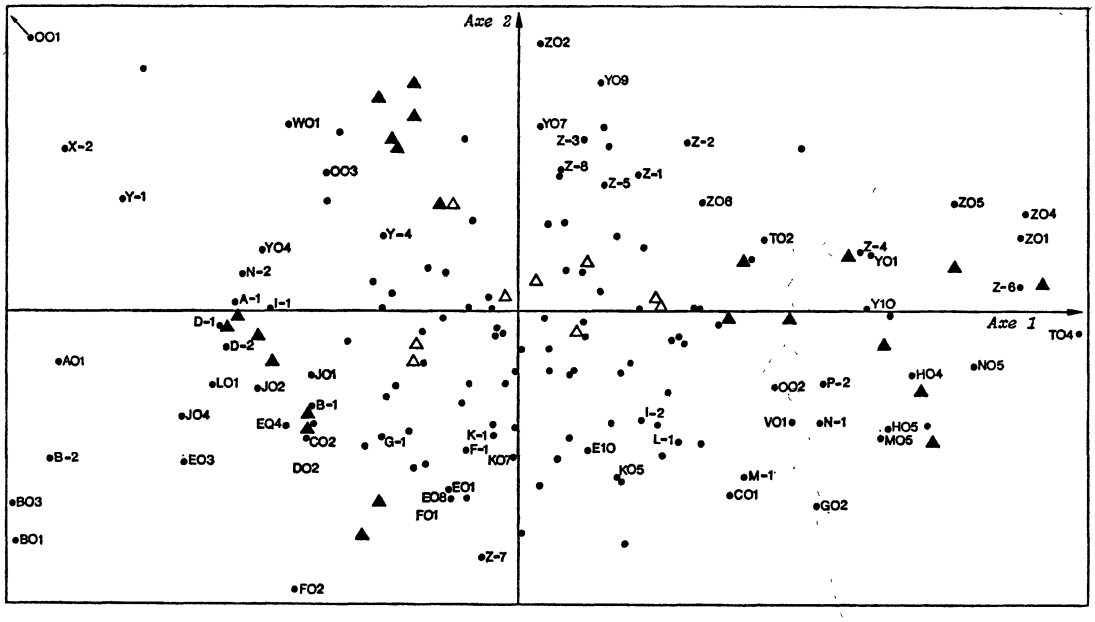
L'analyse a été effectuée dans l'espace R^{32} (de dimension bien plus petite que celle de R^{141})[1]. On verra (1.3) comment à partir des résultats de l'analyse du nuage des V.T.L., $\mathcal{N}(I) \subset R^{32}$, il est possible d'en déduire ceux du nuage des classes d'âges, $\mathcal{N}(J) \subset R^{141}$.

Le tableau 3 fournit les 10 premières valeurs propres et les pourcentages associés de dispersion prise en compte, le tableau 4 les résultats complets relatifs au plus petit ensemble (classes d'âges) sur les 4 premiers axes factoriels : projections, qualité de représentation, et part de chacune des classes d'âge dans la dispersion de chacune des composantes principales définies en 5.2.4.

Nous ne présentons aucun des résultats analogues concernant les V.T.L. car le nombre de tableaux et graphiques est déjà important.

1. Représentation dans le plan des deux premiers axes factoriels.

Figure 1

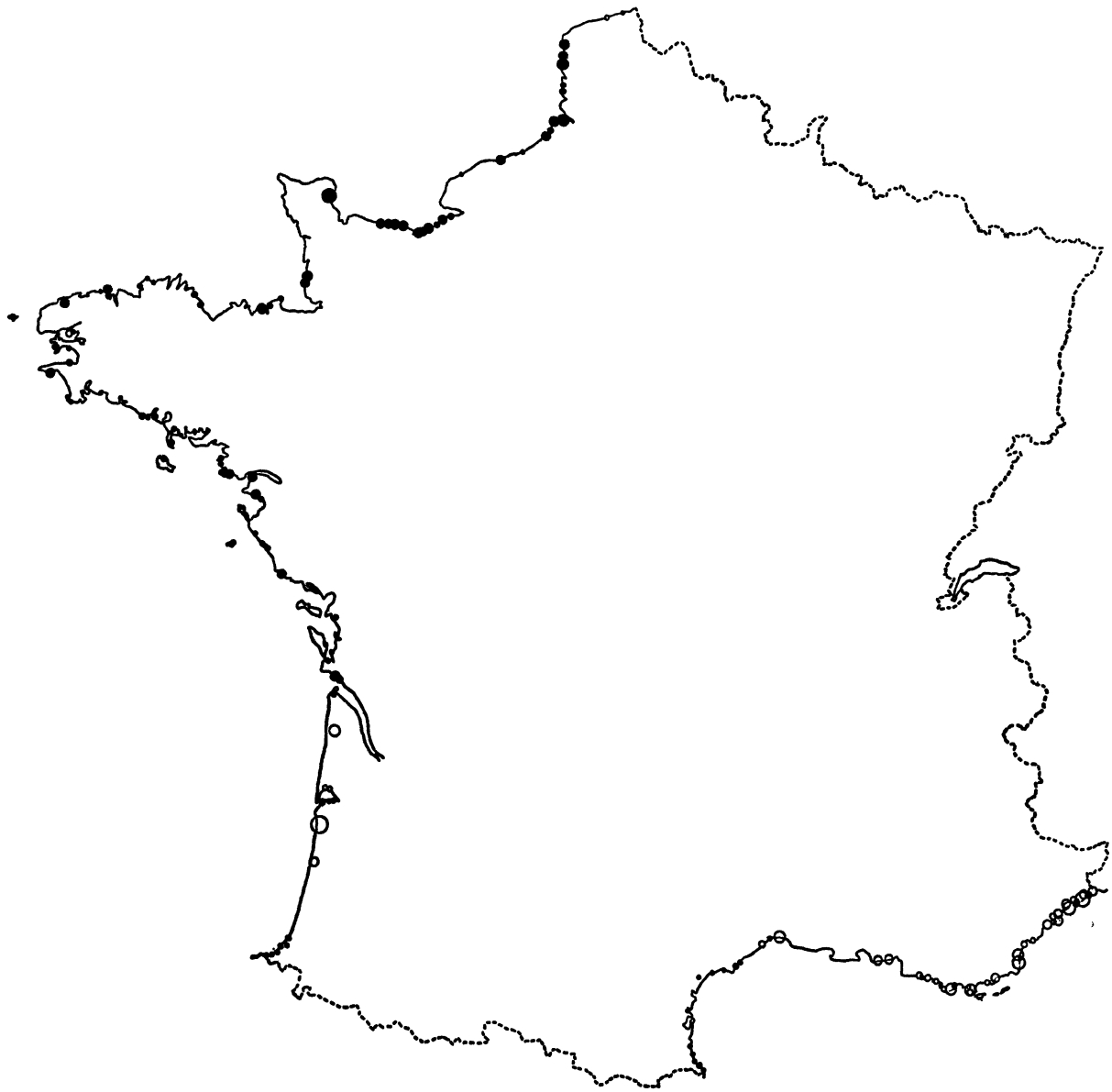


i) Les classes d'âges sont figurées par des triangles ; celles pour lesquelles la qualité de la représentation, par le plan des deux premiers axes factoriels, est supérieure à 0,60 sont figurées par des triangles pleins. Les colonnes 1a et 2a du tableau 4 permettent de nommer les différents triangles.

ii) Les V.T.L. sont figurées par des points ; celles dont la qualité de la représentation, par le plan des deux premiers axes factoriels, est supérieure à 0,60 sont munies de leur code.

1. Le coût des calculs croît avec le cardinal de l'ensemble de plus petit cardinal.

Figure 2. — *Cartographie des villes touristiques littorales sur le 2e axe factoriel. (Les coordonnées positives sont représentées par des cercles vides, les négatives par des cercles pleins. Les diamètres des cercles sont proportionnels aux valeurs absolues de ces coordonnées)*



A l'axe 1 est associée près de 60 % de la dispersion totale. Un effet massif d'opposition jeunes-vieux se manifeste sur cet axe ; les classes d'âge se répartissent (des plus jeunes au plus vieilles) pratiquement dans l'ordre croissant de gauche à droite.

En ce qui concerne les V.T.L. on retrouve une opposition nette entre d'une part les régions touristiques dans lesquelles la part relative des jeunes est importante, le Nord de la France (ex. : le Touquet, Wimereux, Saint-Valéry-en-Caux) ou sur les littoraux plus récemment ouverts au tourisme (Bretagne du Nord), et d'autre part les régions dans lesquelles la proportion des personnes âgées est grande : Côte d'Azur, Côte Vermeille (ex. : Menton, Collioure, etc.).

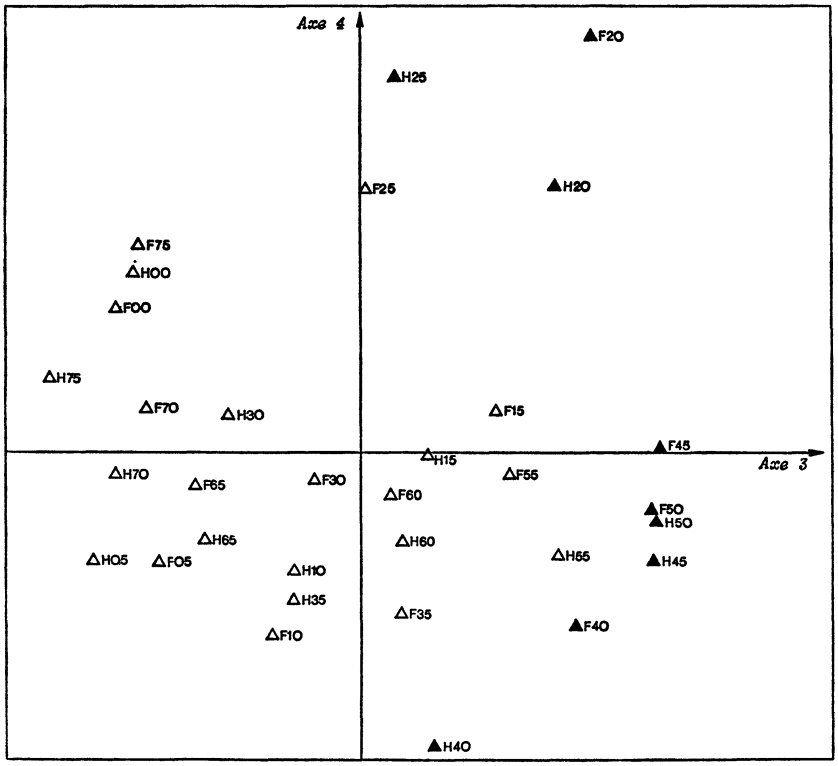
Les résultats obtenus sur l'axe 1 étaient attendus.

L'axe 2 est plus intéressant car il oppose un littoral du Nord et la Bretagne, qui ont un excédent d'adolescents, au littoral méditerranéen qui présente un excédent de jeunes adultes (25-30 ans). La zone Aquitaine-Languedoc constitue une belle transition.

La carte de la figure 2 fournit une illustration de ce résultat. Elle a été constituée à l'aide des projections des descriptions des villes sur le 2e axe factoriel (factor scores) de l'analyse. Les coordonnées positives (resp. négatives) sont représentées par des cercles vides (resp. pleins) dont les diamètres sont proportionnels aux valeurs absolues de ces coordonnées.

Les classes d'âges dont la qualité de la représentation est supérieure à 0,40, par le plan engendré par les 3e et 4e axes factoriels, sont figurées par des triangles pleins.

Figure 3



2. Représentation dans le plan des 3e et 4e axes factoriels

Les 3e et 4e valeurs propres sont très voisines, la part de la dispersion prise en compte par le couple d'axes associés est supérieure à 11 % donc non négligeable. La quasi-égalité de ces valeurs propres est confirmée par d'autres analyses portant sur des données très voisines : analyse des seules V.T.L. de plus de 10 000 habitants, études mettant en jeu les ménages extraordinaires, etc. Pour cette raison le plan (3,4) est considéré comme un plan propre, et pour faciliter l'interprétation des résultats une rotation orthogonale du système d'axes a été effectuée afin d'obtenir des cartographies intéressantes sur ces nouveaux axes. Ces résultats se trouvent dans [13] et nous conseillons au lecteur de s'y reporter. D'autre part, la figure 3 fournit la projection des classes d'âge dans le repère initial et il est assez facile d'en déduire le nouveau repère choisi. Le critère de rotation est simple : l'axe 3′ traverse le groupe des classes d'âge des hommes de 40 à 60 ans et des femmes de 35 à 55 ans, l'axe 4′ traverse les groupes 20-25 ans. (On pourra utiliser le tableau 4 pour détecter les classes d'âges dont la participation à la dispersion du nuage projeté sur le plan (3, 4) est importante).

2. EXPOSÉ DE LA MÉTHODE

2.1. Définitions

2.1.1. *Correspondance*

i) Soit I, J deux ensembles [1] et K un ensemble fini [2] de cardinal N. Soit $P : K \to I \times J$ le protocole d'expérience ou d'enquête initial et $\nu : I \times J \to N$ la distribution d'effectif associée, i.e. l'application définie par :

$$\forall\ (i, j) \in I \times J : n_{ij} = \nu(i, j) = \left|P^{-1}(i, j)\right|$$

l'entier n_{ij} nombre d'occurrences du couple (i, j) est appelé effectif de (i, j).

Exemple (2.1.1.)

K est l'ensemble des habitants des 141 villes littorales françaises,

J : un ensemble de 32 classes d'âges (en fait $J = C_{16} \times S_2$; C = classes d'âge, S = Sexe),

I : l'ensemble des 141 villes touristiques littorales françaises étudiées.

P associe à tout individu le couple constitué par sa classe d'âge et sa ville d'origine.

n_{ij} est l'effectif de la classe j de la ville i.

ii) On note : $n_i = \sum_j n_{ij}$ et $n_j = \sum_i n_{ij}$ les effectifs respectifs de l'élément $i \in I$ et de l'élément $j \in I$; enfin :

$$N = \sum_j n_j = \sum_i n_i = \sum_{i,j} n_{ij} \text{ est l'effectif total}$$

iii) L'application ν est appelée *correspondance* sur $I \times J$ (on dit aussi correspondance de J vers I ou de I vers J). Elle constitue la donnée de base de la méthode étudiée. Nous allons en déduire d'autres distributions.

2.1.2. *Distribution des fréquences - fréquences conditionnelles*

i) On appelle :

fréquence d'occurrence du couple $(i, j) \in I \times J$, le nombre $\dfrac{n_{ij}}{N}$

fréquence d'occurrence de l'élément $i \in I$, le nombre $m_i = \dfrac{n_i}{N}$

fréquence d'occurrence de l'élément $j \in J$, le nombre $m_j = \dfrac{n_j}{N}$

1. On ne tient pas compte de la structure éventuelle de I ou J (structure ordinale par exemple). En tenir compte soulève des problèmes non tous résolus.

2. La condition $|I|, |J|$ finis n'est pas nécessaire ; si elle n'est pas réalisée on restreint ν à un sous-ensemble fini de $I \times J$. Un exemple de restriction de ν est donné en 2.12 avec la condition $n_i n_j \neq 0$.

On suppose que : $\forall i \in I$, $\forall j \in J$, $n_i n_j \neq 0$.

S'il n'en est pas ainsi on restreint les ensembles I ou J en supprimant les éléments i ou j dont les effectifs sont nuls.

ii) On appelle fréquence conditionnelle de l'élément j parmi les n_i occurrences de i le nombre $\frac{n_{ij}}{n_i}$.

On appelle fréquence conditionnelle de l'élément i parmi les n_j occurrences de j le nombre $\frac{n_{ij}}{n_j}$.

iii) On appelle *profil* de l'élément $i \in I$ (resp. $j \in J$) la distribution conditionnelle $(\frac{n_{ij}}{n_j}; 1 \leqslant j \leqslant p)$, (resp. $(\frac{n_{ij}}{n_i}; 1 \leqslant i \leqslant n)$).

Remarque (2.1.2)

La donnée d'une *correspondance* est *équivalente* à la donnée des fréquences conditionnelles $\frac{n_{ij}}{n_i}$, pour tout $(i, j) \in I \times J$ et des effectifs n_i pour tout $i \in I$.
(On peut bien entendu échanger les rôles de I et J.)

2.1.3. *Nuages. Profils et pondérations. Descriptions et descripteurs*

Nous replaçant dans le cadre vectoriel du chapitre 1.2 construisons les tableaux suivants :

Tableau 5

	R^p	R^n
Base	$\mathcal{B} = \{b_j / 1 < j < p\}$	$\mathcal{A} = \{a^i / 1 < i < n\}$
Nuages (notations)	$\mathcal{N}(I)$	$\mathcal{N}(J)$
Vecteurs du nuage profils (ou description)	$x_i = \sum_j \frac{n_{ij}}{n_i} b_j$	$y^j = \sum_i \frac{n_{ij}}{n_j} a^i$
Nombre de profils	n	p
Pondération des profils	au vecteur x_i est affectée la masse $m_i = \frac{n_i}{N}$	au vecteur y^j est affectée la masse $m_j = \frac{n_j}{N}$

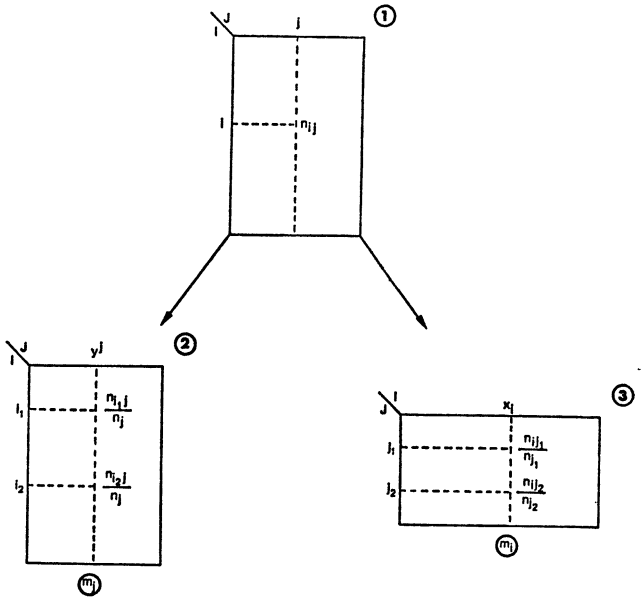
Tableau 6

Descripteurs	J-descripteurs $\sum_i \frac{n_{ij}}{n_i} a^i \in R^n$	I-descripteurs $\sum_j \frac{n_{ij}}{n_j} b_j \in R^p$
Nombre de descripteurs	p	n

On associe à la correspondance, deux nuages :

$$\mathcal{N}(I) = ((x_i ; m_i)/1 \leqslant i \leqslant n) ; \mathcal{N}(J) = ((y^j ; m_j)/1 \leqslant j \leqslant p)$$

Figure 4



(1) tableau associé à la correspondance

(2) tableau associé au nuage $\mathcal{N}(J) = ((y^j, m_j)/j \in J)$

(3) tableau associé au nuage $\mathcal{N}(I) = ((x_i, m_i)/i \in I)$

2.1.4. *Propriétés vectorielles du nuage. Variété linéaire support du nuage*

Considérons dans R^p la variété linéaire $\pi = \{x = \sum_j \alpha_j b_j \in R^p / \sum_j \alpha_j = 1\}$ de dimension $p - 1$.

Soit alors x_i un vecteur du nuage $\mathcal{N}(I) \subset R^p$, on vérifie immédiatement que $x_i \in \pi$ (les coordonnées de x_i possèdent une propriété supplémentaire : elles sont toutes non négatives).

Définition (2.1.4)

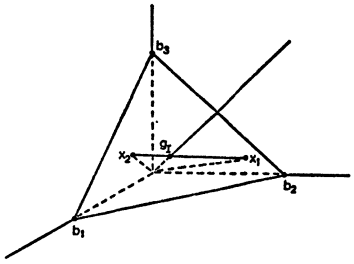
On appelle *support* σ du nuage $\mathcal{N}(I)$ la variété linéaire de plus petite dimension contenant les vecteurs de $\mathcal{N}(I)$.

On remarque que $\sigma \subset \pi$ et $dim\ \sigma \leqslant inf\ (n - 1, p - 1)$; en outre les vecteurs du nuage se trouvent dans la portion $\sigma_T = \{x \in \pi / \forall j \in J : \alpha_j \geqslant 0\}$ domaine polygonal convexe de R^p dont les sommets sont les vecteurs de base b_j de R^p.

Remarque (2.1.4)

Une définition et des résultats analogues peuvent être donnés pour $\mathcal{N}(J)$ dans R^n.

Figure 5



Dans ce cas particulier : $p = 3$, $n = 2$. Alors σ_T est le domaine triangulaire dont les sommets sont b_1, b_2, b_3. La dimension de π est 2, celle de σ est 1.

Propriétés barycentriques

Quel que soit x_i vecteur du nuage $\mathcal{N}(I) \subset R^p$: $x_i = \sum_j \frac{n_{ij}}{n_i} b_j$, on peut exprimer autrement cette propriété en disant que :

x_i est barycentre du nuage $\nu_i = ((b_j, \frac{n_{ij}}{n_i}) / j \in J)$ dont les vecteurs sont ceux de la base $\mathcal{B}$ de R^p.

Pour $i \neq i'$ les vecteurs de ν_i et $\nu_{i'}$ sont évidemment les mêmes ; les pondérations, elles, sont en général différentes. Cette propriété barycentrique jouera un rôle prépondérant dans l'interprétation des résultats (2.4.2.).

Point moyen du nuage $\mathcal{N}(I)$ (voir aussi (I. 1 1))

On note $g(\mathcal{N}(I))$ ou plus simplement g_I le point moyen du nuage $\mathcal{N}(I)$, il est défini par :

$$g_I = \sum_i \frac{n_i}{N} x_i$$

on peut aussi l'exprimer dans la base $\mathcal{B}$,

$$g_I = \sum_j \frac{n_j}{N} b_j = \sum_j m_j b_j$$

Enfin : $g_I \in \sigma$

2.1.5. *Distance euclidienne sur R^p (resp. sur R^n)*

Alors qu'en composantes principales (voir I.2.3.1) la distance euclidienne introduite est définie sans référence [1] aux données, en correspondance, la distance dépend des fréquences d'occurrence des divers éléments.

Produit scalaire ψ : définition

$$\alpha = \sum_j \alpha_j b_j \in R^p,$$

$$\psi(\alpha, \beta) = \sum_j \frac{\alpha_j \beta_j}{m_j},$$

$$\beta = \sum_j \beta_j b_j \in R^p,$$

1. Il n'en est plus ainsi lorsque l'on travaille sur des données centrées et réduites, où la distance est alors une fonction des éléments du tableau de données.

la matrice associée à ψ, relativement à la base $\mathcal{B}$ de $\mathbf{R}^p$, que nous continuerons à noter ψ est :

$$\psi = \begin{pmatrix} m_1^{-1} & & & \\ & \ddots & 0 & \\ & & m_j^{-1} & \\ & 0 & & \ddots \\ & & & & m_p^{-1} \end{pmatrix}$$

Propriétés : outre les propriétés usuelles des distances euclidiennes on a :

a) La base $\mathcal{B}$ de $\mathbf{R}^p$ est ψ -orthogonale mais non ψ -orthonormale, car $||b_j||^2 = \frac{1}{m_j}$ en général différent de 1. La norme de b_j est d'autant plus grande que l'élément $j \in J$ est moins fréquent (m_j petit). Cette particularité est importante car la structure métrique de $\mathbf{R}^p$ en découle : les coordonnées des profils associés aux m_j les plus petits jouent un rôle prépondérant dans le calcul de la norme d'un profil ou dans le calcul de la distance $d^2_\psi(\alpha, \beta) = \sum_j \frac{(\alpha_j - \beta_j)^2}{m_j}$ entre deux profils.

(Il peut arriver que certains éléments soient 100 ou 1 000 fois moins fréquents que d'autres. De telles hétérogénéités poseront des problèmes au moment de la détermination et de l'interprétation des axes factoriels.)

b) Le vecteur moyen g_I est de ψ-norme égale à 1 ; c'est aussi la projection orthogonale du vecteur nul 0 sur le support de $\mathcal{N}(I)$, (fig. 5). En effet :

$$||g_I||^2_\psi = \sum_j [m_j^2 \, ||b_j||^2_\psi] = 1$$

quel que soit $i \in I$: $\psi(x_i, g_I) = \sum_j \frac{m_j \frac{n_{ij}}{n_i}}{m_j} = 1$

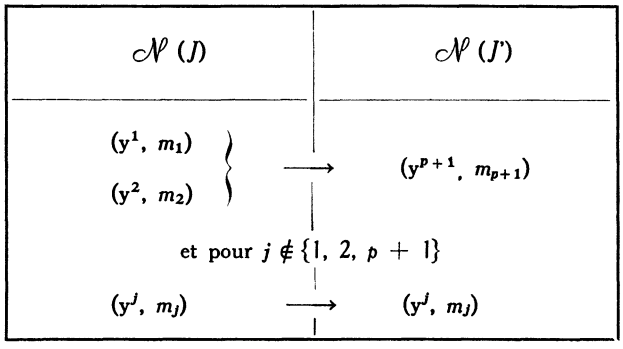
d'où $\psi(x_i, g_I) = ||g_I||_\psi$. Donc l'axe $[g_I]$ est ψ-orthogonal au support du nuage $\mathcal{N}(I)$; en conséquence g_I est bien la projection ψ-orthogonale de 0 sur le support du nuage $\mathcal{N}(I)$.

c) *Remarques sur distances et profils*

1. La métrique définie est parfois appelée métrique du χ^2. Cette terminologie rappelle sa parenté (ou l'analogie) avec la statistique du même nom $\chi^2 = \sum_j \frac{[n_j - Np_j]^2}{Np_j}$ les notations étant traditionnelles. Cette quantité est proportionnelle au carré de la distance (associée au produit scalaire sur $\mathbf{R}^p$ défini par : $\sum_j \frac{\alpha_j \beta_j}{p_j}$ entre une distribution de fréquence $(\frac{n_j}{N}, 1 \leqslant j \leqslant p)$ et la distribution de probabilité de l'hypothèse nulle $(p_j, 1 \leqslant j \leqslant p)$).

2. Comme on l'a vu la définition des vecteurs du nuage (2.1.3) amène à confondre (vecteurs à distance nulle) les descriptions d'éléments i et i' pour lesquels les distributions d'effectifs sont proportionnelles et non pas seulement égales. Deux telles descriptions sont dans une même classe d'équivalence pour une relation d'équivalence dite distributionnelle; chaque élément est en effet décrit par la distribution de ses occurrences. Le terme d'équivalence distributionnelle renvoie à une définition du linguiste Z.S. Harris ; pour ce point d'histoire on peut se reporter à Benzécri (11b, p. 189).

3. Supposons $y^1 = y^2$. Posons $y^{p+1} = y^1$ et $m_{p+1} = m_1 + m_2$.
Soit $\mathcal{N}(J') = ((y^j, m_j)/3 \leqslant j \leqslant p + 1)$. On a :



Soit $\mathcal{N}'(I) \subset \mathbf{R}^{p-1}$ le nuage associé à cette nouvelle correspondance. Soit Φ' et ψ' les métriques définies sur $\mathbf{R}^n$ et $\mathbf{R}^{p-1}$ comme en 2.1.5. On montre facilement que :

i) dans $\mathbf{R}^n$
- $\Phi = \Phi'$,
- les distances entre vecteurs homologues de $\mathcal{N}(J)$ et $\mathcal{N}(J')$ sont égales,
- quel que soit P sous-espace vectoriel *Disp* $\mathcal{N}_p(J')$ = *Disp* $\mathcal{N}_p(J)$.

ii) dans $\mathbf{R}^{p-1}$
- $\forall k,\ 1 \leqslant k, l \leqslant n - 1 \quad ||x'_k - x'_l||_{\psi'} = ||x_k - x_l||_{\psi}$

x'_k et x'_l étant les vecteurs de $\mathcal{N}'(I)$ associés aux éléments k et l.

Pour plus de détails sur cette question, on peut se reporter à [12].

Ces propriétés sont l'analogue des propriétés que l'on peut établir sans difficulté dans le cadre général du § I.2 pour des nuages initiaux :

$\mathcal{N}(I) = ((x_i, \frac{1}{n}),\ 1 \leqslant i \leqslant n)$ de $\mathbf{R}^p$, la métrique étant $\psi = (m_j\, \delta_{jj'})$.

$\mathcal{N}(J) = ((y^j, m_j),\ 1 \leqslant j \leqslant p)$ de $\mathbf{R}^n$, la métrique étant $\Phi = (\frac{1}{n}\, \delta_{ii'})$.

L'importance de la propriété de la Remarque 3 n'est pas complètement explorée. Cette propriété n'est jamais directement utilisée ; mais c'est sans doute en s'appuyant sur elle que l'on peut penser que certains résultats de l'analyse d'un tableau de grandes dimensions sont peu modifiés (et donc *stables*) lorsqu'on regroupe des éléments.

2.2. Meilleurs ajustements par un sous-espace de dimension r

On a vu (I.2.4) que la résolution de ce problème nécessite la détermination des vecteurs et valeurs propres d'une matrice facilement déterminée à partir de la forme quadratique de dispersion.

Dans ce qui suit, étudiant dans $\mathbf{R}^p$ le nuage $\mathcal{N}(I)$ nous établissons la matrice Σ puis la matrice à diagonaliser. Des remarques facilitant la recherche des axes principaux ψ-orthogonaux achèvent ce paragraphe.

2.2.1. *Notations matricielles*

Soit F la matrice $n \times p$: $F = (n_{ij})$ $1 \leqslant i \leqslant n$, $1 \leqslant j \leqslant p$

Φ la matrice $n \times n$: $\Phi = (m_i^{-1}\, \delta_{ii'})$,

ψ la matrice $p \times p$: $\psi = (m_j^{-1}\, \delta_{jj'})$.

$\delta_n = {}^t(1\ 1\ 1\ \ldots\ 1) \in \mathbf{R}^n$ et $\delta_p = {}^t(1\ 1\ \ldots\ 1) \in \mathbf{R}^p$

Au nuage $\mathcal{N}(J) \subset \mathbf{R}^n$ est associée la matrice $n \times p$, $B = F\psi$ dont la j^e colonne est la suite des coordonnées de y^j.

Au nuage $\mathcal{N}(I) \subset \mathbf{R}^p$ est associée la matrice $p \times n$, $A = {}^tF\Phi$ dont la i^e colonne est la suite des coordonnées de x_i.

La matrice colonne $p \times 1$ associée à g_I sera notée G ;

La matrice colonne $p \times 1$ associée à g_J sera notée K.

D'après (I.2.4), Σ s'exprime de la façon suivante :

$\Sigma = \psi\,{}^tF\Phi F\psi - \psi G\,{}^tG\psi = L - \psi G\,{}^tG\psi = L - \delta_p\,{}^t\delta_p = \psi AB - \delta_p\,{}^t\delta_p$; son terme général est $s_{jj'} = Cov_\Phi(y^j, y^{j'}) = \Phi(y^j, y^{j'}) - 1$, et le terme général de $L = \psi AB$ est $l_{jj'} = \Phi(y^j, y^{j'})$.

2.2.2. *Recherche de la solution*

Il reste à déterminer, conformément aux résultats du (§ I.2.) r vecteurs propres u_k, ψ-orthonormés associés aux r plus grandes valeurs propres de la matrice :

$$\psi^{-1}\Sigma = {}^tF\Phi F\psi - G\,{}^tG\psi = \psi^{-1}L - G\,{}^tG\psi.$$

Nous commençons par montrer que la diagonalisation de cette matrice *non-symétrique* peut-être obtenue à l'aide de celle d'une matrice symétrique pour laquelle existent des algorithmes et des programmes très efficaces. Les propriétés qui suivent sont introduites dans ce but.

(P_1) - *Comparaisons des vecteurs propres de* $\psi^{-1}\Sigma$ *et de* $\psi^{-1}L$

a) G est vecteur propre de $\psi^{-1}\Sigma$ associé à la valeur propre 0 et de $\psi^{-1}L$ associé à la valeur propre 1,

b) Tout autre vecteur propre (resp. valeur propre) de $\psi^{-1}\Sigma$ est vecteur propre (resp. valeur propre) de $\psi^{-1}L = AB$.

Preuve :

a) La démonstration, immédiate, est laissée au lecteur.

b) Soit U un vecteur colonne ψ-orthogonal à G : ${}^tG\psi U = 0$.

$$\psi^{-1}\Sigma U = \psi^{-1}LU - G\,{}^tG\psi U = \psi^{-1}LU,$$

donc : $(\psi^{-1}\Sigma)\ \lambda U = \lambda U \Leftrightarrow (\psi^{-1}L)\ U = \lambda U$, ce qui achève la démonstration.

Remarquons que les droites vectorielles [u] de $\mathbf{R}^p$ engendrées par les vecteurs colonnes U sont parallèles à la variété linéaire π de dimension $p - 1$; de plus si $\lambda \neq 0$, [u] est parallèle au support du nuage.

(P_2) - U est vecteur propre de $\psi^{-1}L$ associé à la valeur propre λ si et seulement si $U' = \psi^{1/2}U$ est un vecteur propre de la *matrice symétrique* positive $\psi^{-1/2}L\psi^{-1/2}$ associé à la même valeur propre, en notant

$$\psi^{1/2} = \left(\frac{1}{\sqrt{m_j}}\,\delta_{jj'}\right) \text{ et } \psi^{-1/2} = \left(\sqrt{m_j}\,\delta_{jj'}\right)$$

Preuve : Soit (U, λ) tel que : $\psi^{-1}LU = \lambda U$

on a $\psi^{-1}L\psi^{-1/2}U' = \lambda\psi^{-1/2}U'$

et en multipliant les deux membres par $\psi^{1/2}$: $\psi^{-1/2}L\psi^{-1/2}\ U' = \lambda U'$.

Conséquences : α. Les valeurs propres sont réelles non négatives : $0 \leqslant \lambda_k$

β. On détermine r valeurs propres [1] de la matrice symétrique $\psi^{-1/2}L\psi^{-1/2}$, les vecteurs cherchés $(U_k)_{1\leqslant k\leqslant r}$ se déduisent des vecteurs $(U'_k)_{1\leqslant k\leqslant r}$ par la relation $U_k = \psi^{-1/2} U'_k$. Ils sont ψ-orthonormés car : ${}^tU_k\psi U_k = {}^tU'_kU'_k = \delta_{kk'}$.

2.2.3. *Propriétés des valeurs propres*

La matrice $\psi^{-1}L = (\psi{}^tF)(F\psi) = (\gamma_{jj'})_{1\leqslant j'j\leqslant p}$ est stochastique [19] i.e. telle que pour tout $j \in]p]$: $\Sigma_{j'}\gamma_{jj'} = 1$. Comme toute matrice stochastique, ses valeurs propres sont inférieures à un en module. Donc d'après 2.2.2.d :

$$\forall k \leqslant n : 0 \leqslant \lambda_k \leqslant 1$$

2.2.4. *Axes principaux et composantes principales* (voir définition (3.3.1)).

Les axes $([u_k])_{1\leqslant k\leqslant r}$ sont les *axes principaux* du nuage (ou axes factoriels).

La dispersion du nuage projeté sur $[u_k]$ est λ_k, k^e valeur propre de $\psi^{-1}\Sigma$ ou $\psi^{-1}L$; on n'indexe ni [G] ni la valeur propre qui lui est associée [2].

Posons $x_{ik} = \psi(x_i, u_k)$, alors $c^k = (x_{1k}, ..., x_{nk})$ est la k^e *composante principale* du nuage $\mathcal{N}(I) \subset R^p$. c^k est la suite des n coordonnées des projections des vecteurs du nuage $\mathcal{N}(I)$ sur u_k.

2.3. Etude conjointe des deux nuages $\mathcal{N}$ (I) et $\mathcal{N}$ (J)

Une étude semblable à celle faite pour $\mathcal{N}(I) \subset R^p$ pourrait être faite pour $\mathcal{N}(J) \subset R^n$. Il est alors facile d'établir la matrice associée à la forme quadratique de dispersion du nuage $\mathcal{N}(J)$ ainsi que de déterminer les axes principaux.

Inscrivons sous forme de tableau les résultats relatifs aux deux nuages.

Tableau 7

R^p	R^n
$\mathcal{N}(I)$	$\mathcal{N}(J)$
$\Sigma = \psi AB - \delta_p {}^t\delta_p$ Σ est une matrice $p \times p$	$\mathcal{C} = \Phi BA - \delta_n {}^t\delta_n$ $\mathcal{C}$ est une matrice $n \times n$
$s_{jj'} = Cov\psi(y^j, y^{j'})$	$t_{ii'} = Cov\Phi(x_i, x_{i'})$
Les vecteurs propres u_k sont ceux de AB. $\mathcal{U} = \{u_k / 1 < k < p\}$ base ψ-orthonormale	Les vecteurs propres v_k sont ceux de BA. $\mathcal{V} = \{v^h / 1 < h < n\}$ base Φ-orthonormale
k^e composante principale $c^k = (x_{1k}, ..., x_{nk})$ avec $x_{ik} = \psi(x_i, u_k)$	h^e composante principale $c_h = (y^{1h}, ..., y^{ph})$ avec $y^{jh} = \Phi(y^j, v^h)$

1. On continue à supposer les valeurs propres distinctes (voir note 1 du § I. 2.4).

2. Remarquons qu'à [G] correspond la plus grande valeur propre, égale à 1, de $\psi^{-1}L$ et la plus petite égale à 0 de $\psi^{-1}\Sigma$.

2.3.1. *Relations liant les axes principaux de* $\mathcal{N}$ (I) *et* $\mathcal{N}$ (J)

Soit α et β les applications dont les matrices sont A et B dans le couple de base $\mathcal{A}$, $\mathcal{B}$. Les deux familles de valeurs propres non nulles des applications $\alpha\circ\beta$ et $\beta\circ\alpha$ sont égales [21a] et les sous-espaces propres de $\alpha\circ\beta$ et $\beta\circ\alpha$ associés à une même valeur propre non nulle sont image l'un de l'autre respectivement par β et α.

Ainsi si $\lambda_k \neq 0$ $(\lambda_k, [u_k])$ étant associé à $\mathcal{N}$ (I) et $(\lambda_k, [v^k])$ associé à $\mathcal{N}$ (J)

β $([u_k]) = [v^k]$ et α $([v^k]) = [u_k]$;

plus précisément, si $||u_k||_\psi = 1$ et $||v^k||_\Phi = 1$, alors :

$$\lambda_k \neq 0 : \frac{\beta[u_k]}{\sqrt{\lambda_k}} = v_k \text{ et } \frac{\alpha[v_k]}{\sqrt{\lambda_k}} = u_k$$

Remarque (2.3.)

Il est intéressant d'étudier en détail les applications α et β.

a) Les images par β des vecteurs de la base $\mathcal{B}$ sont les profils y^j ;
Les images par α des vecteurs de la base $\mathcal{A}$ sont les profils x_i.

b) Dans le couple de base $\mathcal{U}$, $\mathcal{V}$ les matrices associées à α et β sont diagonales :

$$Mat\ (\alpha, \mathcal{U}, \mathcal{V}) = \begin{array}{|c|c|}\hline \sqrt{\lambda_k} & 0 \\ \hline 0 & 0 \\ \hline \end{array} \quad \text{avec } \lambda_k = 0 \text{ si } k > rang\ (AB)$$

$$Mat\ (\alpha, \mathcal{U}, \mathcal{V}) = {}^t Mat\ (\beta, \mathcal{U}, \mathcal{V})$$

c) Notons que $(u_k / k \in]p])$ et $(v^k / h \in]n])$ sont des vecteurs propres à droite et à gauche (orthogonaux respectivement pour les métriques Φ et ψ) de la matrice rectangulaire B. Pour plus de détails concernant la recherche des vecteurs propres de matrices rectangulaires se reporter à [21b]. Pour une étude plus systématique des applications α et β et l'intégration des méthodes factorielles dans un cadre plus général voir [14].

2.4. Représentations graphiques et remarques pour leur utilisation

2.4.1. *Principes à partir desquels sont effectuées les représentations graphiques*

a) Les représentations sont :

. axiales ou planes,

.. obtenues par projection Φ-orthogonale dans R^n (ou ψ-orthogonale dans R^p)

... constituées par les projections des éléments des nuages et parfois des vecteurs de base.

.... Dans R^n (par exemple) la métrique représentée est Φ : le cos_Φ de deux vecteurs de R^n est représenté par un cosinus associé au produit scalaire usuel (identité) de R^n. Les longueurs mesurées en centimètres des vecteurs du plan sont proportionnelles à leur Φ-norme.

b) Parmi les graphiques qui peuvent être construits, trois sont courants, le plus habituel, le troisième, n'est pas le plus simple à interpréter.

Les notations étant celles de 2.2.4 on a :

	Graphique 1 : R^p	Graphique 2 : R^n
Métrique représentée	ψ	Φ
Projection sur le plan	$[u_{k'}, u_k = P]$	$[v^k, v^{k'}] = Q$
Vecteurs de base (peuvent être omis)	$1 < j < p \quad Pr_P(b_j) = b_{jk}u_k + b_{jk'}u_{k'}$	$1 < i < n \quad Pr_Q(a^i) = a^{ik}v^k + a^{ik'}v^{k'}$
Profils	$1 < i < n \quad Pr_P(x_i) = x_{ik}u_k + x_{ik'}u_{k'}$	$1 < j < p \quad Pr_Q(y^j) = y^{jk}v^k + y^{jk'}v^{k'}$

Dans le *graphique* 3, représentation dite simultanée : les métriques représentées sont à la fois Φ et ψ. On regroupe en un même graphique les projections $Pr_P(x_i)$ soit n vecteurs et $Pr_Q(y^j)$ soit p vecteurs. Les projections des vecteurs de base sont omises.

c) Exemple de l'étude des villes touristiques littorales

Nous avons introduit deux représentations graphiques :

i) la figure 1 qui est un graphique de type 3 (représentation dite simultanée) ;

ii) la figure 3 qui est un graphique de type 2 (sans vecteurs de base).

2.4.2. *Trois propriétés utiles à l'interprétation*

(P'1) - D'après la propriété barycentrique de 1.1.4 on a :

sur le graphique 1 : $Pr_P(x_i)$ est le point moyen du nuage $(Pr_P(b_j) ; \frac{n_{ij}}{n_i} \;/\; 1 \leqslant j \leqslant p)$

sur le graphique 2 : $Pr_Q(y^j)$ est le point moyen du nuage $(Pr_Q(a^i) ; \frac{n_{ij}}{n_j} \;/ 1 \leqslant i \leqslant n)$

(P'2) - *Produits d'affinités*

Soit $s = dim\ \sigma$ où σ, on le rappelle, est le support du nuage $\mathcal{N}(I)$, puisque $a^i = \sum_{k=1}^{n} a^{ik}v^k$ (expression des vecteurs de $\mathcal{A}$ dans la base $\mathcal{V}$) d'après 3.3.2. :

$$x_i = \alpha(a^i) = \alpha\left(\sum_{k=1}^{n} a^{ik}\, v^k\right) = \sum_{k=1}^{n} a^{ik}\, \alpha(v^k)$$

$$= \sum_{k=1}^{s} a^{ik}\sqrt{\lambda_k}\, u_k$$

(la dimension de σ étant s, seules les s plus grandes valeurs propres de $\psi^{-1}\Sigma$ sont non nulles).

$$= \sum_{k=1}^{s} x_{ik}\, u_k$$

donc $\forall k \in]s]$, $\boxed{x_{ik} = \sqrt{\lambda_k}\, a^{ik} = \psi(x_i, u_k)}$ (1)

Quel que soit $i \in]n]$:

— x_i s'écrit $(x_{i1}, ..., x_{is}, 0, ..., 0) \in R^p$ dans la base $\mathcal{U}$.

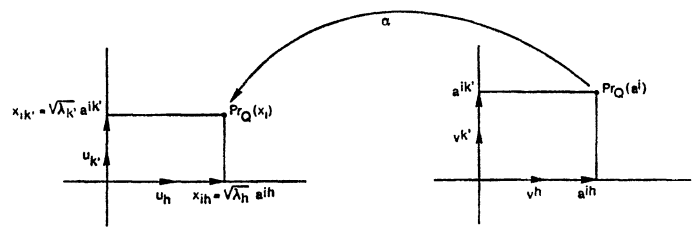
— la relation (1) permet de dire que le vecteur $(x_{i1}, ..., x_{is}) \in R^s$ est l'image de $(a^{i1}, ..., a^{is}) \in R^s$ par la composée de s affinités de rapports $\sqrt{\lambda_k}$, $(1 \leqslant k \leqslant s)$ et d'axes directeurs ceux engendrés par les vecteurs de la base canonique de R^s.

— matriciellement on peut écrire dans la base canonique de R^s :

$$\begin{pmatrix} x_{i1} \\ \cdot \\ \cdot \\ \cdot \\ x_{is} \end{pmatrix} = \begin{pmatrix} \sqrt{\lambda_1} & & \\ & \ddots & 0 \\ & 0 & \ddots \\ & & & \sqrt{\lambda_s} \end{pmatrix} \begin{pmatrix} a^{i1} \\ \cdot \\ \cdot \\ \cdot \\ a^{is} \end{pmatrix}$$

de même : $\forall k \in]s]$ $\boxed{y^{jk} = \sqrt{\lambda_k}\ b_{jk}}$

Figure 6



(P'3) - *Relation barycentrique et produit d'affinité*

De (P'1) et (P'2) on déduit que :

$$\forall i \in]n] \quad \forall k \in]s] \quad : x_{ik} = \frac{1}{\sqrt{\lambda_k}} \sum_{j=1}^{p} \frac{n_{ij}}{n_i}\ y^{jk}$$

2.4.3. *Remarque pour l'utilisation des graphiques*

a) Il s'agit toujours à l'aide de procédures inférentielles basées sur des constatations faites sur les graphiques ou les tableaux de résultats de déterminer des propriétés sur les liaisons entre les $i \in I$, ou entre les $j \in J$, ou entre les éléments de I et ceux de J.

Les éléments sont représentés sur les différents graphiques par les projections des vecteurs suivants :

Graphiques	1	2	3
Eléments i	a^i	b_j	x_i
Eléments j	y^j	x_i	y^j

b) Afin de juger, à l'aide des graphiques et tableaux, des liaisons entre profils (proximités de profils ou de profils et de vecteurs de base) on dispose pour chacun des éléments projetés de sa qualité de représentation par le plan de figure. Cet indice est utilisé comme en composantes principales.

Sur les graphiques 1, 2 et 3 (pour deux profils x_i ou deux profils y^j) deux vecteurs bien représentés et de projections proches sont proches.

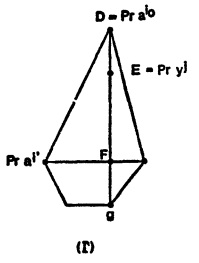
Exemples V.T.L. : la figure VI par exemple permet de mettre en évidence des groupements de villes bien représentés aux deux extrémités gauche (Nord : BRAY-DUNES, LE PORTEL, WIMEREUX, EQUIHEN) et droite (Méditerranée : BEAULIEU, ROQUEBRUNE, SAINT-JEAN-CAP-FERRAT, COLLIOURE) ; il en est de même pour les classes d'âges que le lecteur pourra regarder.

c) *Vecteurs mal représentés*

i) Si la projection d'un profil y^j est proche, par exemple, de celle d'un vecteur de base a^{i_o}, peut-on en déduire que $\frac{n_{i_o j}}{n_j}$ est grand?

- Oui si $Pr\ (a^{i_o})$ est périphérique (voir la figure 6 bis).
- Non si $Pr\ (a^{i_o})$ n'est pas un vecteur de la périphérie du graphique comme on peut s'en convaincre sur un grand nombre d'exemples.

Figure 6 bis



En effet $Pr\ y^j$ est barycentre de $Pr\ a^{i_o}$ muni de la masse $\frac{n_{i_o j}}{n_j}$ et du point moyen g du nuage $\left\{ (Pr\ (a^i), \frac{n_{ij}}{n_j})\ ;\ i \neq i_o,\ 1 \leqslant i \leqslant n \right\}$ affecté de la masse $1 - \frac{n_{i_o j}}{n_j}$.

Soit (Γ) l'enveloppe convexe de $(Pr\ (a^i),\ i \neq i_o)$ g est à l'intérieur de (Γ). Les points D, E, g sont alignés. F étant définie sur la figure (6 bis) on a :

$$1 \geqslant \frac{n_{i_o j}}{n_j} \geqslant \frac{\| E - F \|}{\| D - g \|} \geqslant \frac{\| E - F \|}{\| D - F \|}, \text{ et si } \| E - F \| \simeq \| D - F \|$$

$$\text{alors } \frac{n_{i_o j}}{n_j} \simeq 1.$$

ii) Il peut arriver que $Pr\ y^j$ soit proche d'un sous-ensemble, $\{Pr\ (a^i)\ /\ i \in I_o \subset I\}$, de vecteurs situés à la périphérie du graphique. Dans ce cas, on peut déduire de cette proximité :

$$\sum_{i \in I_o} \frac{n_{ij}}{n_i} \simeq 1.$$

d) Enfin dans le graphique 3 les relations liant les x_{ik} aux y^{jk} faisant intervenir des affinités (voir (P2) et (P3) de 1.4.2) sont en général suffisamment compliquées pour ne pas autoriser l'utilisation des propriétés barycentriques avec la même sécurité que dans les graphiques 1 et 2.

Exemple des V.T.L. : la petite taille des valeurs propres, mais aussi la grande différence entre les deux premières ($\sqrt{\lambda_1} \simeq 0{,}4$, $\sqrt{\lambda_2} \simeq 0{,}06$ toutes deux très différentes de 1) nous interdisent une utilisation aveugle des proximités entre classes d'âges et V.T.L., même si elles sont toutes bien représentées. Cela dit, les données confirment les liaisons entre les villes du sud déjà citées et les classes d'âges représentant les personnes de plus de 70 ans.

2.4.4. *Part de la dispersion prise en compte par les éléments* x_i *et* y^j

Etant donné x_i vecteur du nuage $\mathcal{N}(I)$, $[u_k]$ un axe factoriel, on appelle part de la dispersion de x_i dans la dispersion du nuage projeté sur $[u_k]$ l'expression

$$\Delta_i^k = \frac{m_i}{\lambda_k} \left\| Pr_{[u_k]} (x_i - g_I) \right\|_\psi^2 ; \quad 0 \leqslant \Delta_i^k \leqslant 1.$$

Les différentes colonnes b du tableau 4 fournissent ces éléments en pourcentage pour les classes d'âges et les 4 premiers axes factoriels.

a) Si les m_i sont presque égaux entre eux, pour k fixé les Δ_i^k sont proportionnels aux $\left\| Pr_{[u_k]} (x_i - g_I) \right\|^2$; ces quantités se déduisent des graphiques.

b) Si les m_i sont différents les uns des autres, les valeurs des Δ_i^k ne peuvent être déduites des graphiques.

Dans les deux cas cet indice constitue un complément important au critère de la qualité de la représentation et permet d'évaluer (ou d'apprécier) la contribution des différents éléments (points pondérés) du nuage à la détermination de l'axe.

Remarque (2.4.4)

Une définition et des propriétés analogues peuvent être énoncées pour $y^j \in \mathbb{R}^n$ et $[v^h]$ un axe factoriel du nuage $\mathcal{N}(J)$.

2.4.5. *Représentation d'éléments supplémentaires et suppression d'éléments*

a) Considérons deux correspondances l'une sur $I \times J_1$ l'autre sur $I \times J_2$.

Exemple : Dans l'analyse des compositions par âges des V.T.L. on peut considérer l'ensemble J_1 des classes d'âges hommes, et l'ensemble J_2 des classes d'âges femmes. Les analyses de $I \times J_1$ et $I \times J_2$ fournissent deux ensembles de résultats (graphiques, tableaux, interprétations).

Il est intéressant de rapporter, par exemple, les descripteurs de J_2 aux différentes représentations obtenues sans eux ; cela a pour but de confronter l'analyse faite sur $I \times J_1$ à des informations n'intervenant pas dans l'analyse.

La méthode est simple : il suffit de projeter les descripteurs de J_2 sur les axes factoriels de l'analyse dans l'analyse $I \times J_1$. Les graphiques et les qualités de représentation des vecteurs associés à J_2 ainsi obtenus sont utilisés pour étudier par exemple les liens entre éléments de J_1 et J_2.

b) Considérons un vecteur $x_i \in \mathbb{R}^p$ tel que $m_i \simeq 0$ et dont la part de la dispersion dans la dispersion totale $\left(m_i \|x_i - g_I\|_\psi^2 \Big/ \sum_i m_i \|x_i - g_I\|^2 \right)$ est grande. Ce vecteur peut être déterminant dans le calcul d'un des premiers axes factoriels. On se trouve alors dans une situation parfois embarrassante : un des premiers axes (qui prend en compte une grande part de la dispersion totale du nuage) est en fait spécifique à un seul vecteur. Dans ce cas il est intéressant d'effectuer une analyse sur les ensembles J et $I - \{i\}$ et de projeter le vecteur x_i en élément supplémentaire sur les axes obtenus. Cette pratique revient à affecter à x_i une masse nulle.

1. Cet indice et ces procédures peuvent être introduits pareillement en composantes principales.

3. Remarques et conclusions

3.1. Quelques-uns des liens qui unissent les méthodes des composantes principales (B) et des analyses de tableaux de correspondances (C) ont été développés précédemment. Il est même facile de montrer que : $\mathcal{N}(I)$ étant le nuage défini en 2.1.3, nuage de $(\mathbf{R}^p, \psi)$, la recherche des axes principaux $\{[u_k]\ 1 \leqslant k \leqslant p\}$ par la méthode (C) se ramène à la résolution d'un problème de détermination d'axes principaux par la méthode (B). Plus précisément, il existe dans $\mathbf{R}^p$ muni du produit scalaire i identité :

— un nuage $\mathcal{N}'(I)$;

— une isométrie linéaire l de $(\mathbf{R}^p, \psi)$ sur $(\mathbf{R}^p, \mathrm{i})$, tels que $\{[u'_k] / 1 \leqslant k \leqslant p\}$ étant le système des axes principaux de $\mathcal{N}'(I)$ solution obtenue par la méthode (B) on a :

— $\forall k \in]p] : l(u'_k) = u_R$

— les graphiques associés aux projections pour la métrique i de $\mathcal{N}'(I)$ sur $[u'_k, u'_{k'}]$ par exemple et ceux associés aux projections pour la métrique ψ de $\mathcal{N}(I)$ sur $[u_k, u_{k'}]$ sont identiques ; en outre les qualités de représentation des points homologues sont égales.

L'isométrie linéaire est définie dans la base $\mathcal{B}$ par $l^{-1}(b_j) = \dfrac{b_j}{\sqrt{m_j}}$: c'est un changement d'unité sur les axes b_j.

Si $\mathcal{N}(I) = \{(x_i, m_i) / 1 \leqslant i \leqslant n\}$

$\mathcal{N}'(I) = \{(l^{-1}(x_i), m_i) / 1 \leqslant i \leqslant n\}$,

les vecteurs u_k sont déterminés en 2.2.2;

les vecteurs u'_k sont les vecteurs propres de la matrice Z obtenue en appliquant les résultats du chapitre 1.2.4., c'est-à-dire :

$$Z = \psi^{1/2}\,{}^t F \Phi F\, \psi^{1/2} - \psi^{-1/2}\delta_p\ {}^t\delta_p\ \psi^{-1/2},$$

que l'on comparera à la matrice $\psi^{-1/2}(L - \delta_p\ {}^t\delta_p)\ \psi^{-1/2}$ introduite en 2.2.1 dans la propriété (P2).

3.2. Les notations étant celles des § I.2.1 et I.2.2, dans (E, ψ) associons à un nuage $\mathcal{N}$ une fonction :

$$L(\mathcal{N}) = \sum_i m_i \|x_i\|^2.$$

Cette quantité, qui est *l'inertie* du nuage par rapport à l'origine, est aussi souvent utilisée que la dispersion, à laquelle elle est liée par $Disp(\mathcal{N}) = L(\mathcal{N}) - \|G\|^2$.

Pareillement à I.2.2, on définit $L(\mathcal{N}_P) = \Sigma\, m_i \|Pr_P(x_i)\|^2$, $P \subset E$ étant un sous-espace vectoriel et $\mathcal{N}_P$ le nuage projeté sur P.

On a $L(\mathcal{N}) = L(\mathcal{N}_P) + L(\mathcal{N}_{P\perp})$.

Pareillement à I.2.3, on définit l'ajustement linéaire de la fonction L du nuage par un sous-espace vectoriel de E en remplaçant dans I.2.3.1 le terme *Dispersion* par le terme *Fonction L.*

En général les deux méthodes, l'une (L) basée sur la fonction L [1], l'autre (D) sur la dispersion, déterminent des axes principaux différents.

Cependant G étant le point moyen du nuage si l'une des conditions suivantes est satisfaite [2] :

a) $G = 0$.

b) [G] est orthogonal au support du nuage ou ce qui est équivalent,

[G] est axe principal pour (D) associé à la valeur propre 0.

c) [G] est axe principal du nuage pour (D) et pour (L)

alors les axes principaux pour (D) et pour (L) sont identiques, les qualités d'ajustement par les axes sont les mêmes sauf pour l'axe [G] qui est tel que : $Disp\ \mathcal{N}_{[G]} = L(\mathcal{N}_{[G]}) - \|G\|^2$.

---

1. Pour un exemple voir Kendall et Stuart [20].

2. On a : a ⇒ c et b ⇒ c.

Le cas des correspondances est le cas b : le vecteur moyen est la projection orthogonale de 0 sur σ (support du nuage) ; voir 1.1.5. La recherche des axes (D) se ramène (comme nous l'avons montré en I.3.2.1) à la résolution d'un problème (L). Cette simplification ne se retrouve pas en général dans l'analyse en composantes principales.

3.3. Le cadre du chapitre I.2, élargi par l'introduction de la fonction L définie dans la remarque précédente, permet d'étudier l'ajustement linéaire d'un nuage équi ou non équipondéré par un sous-espace engendré par un système d'axes ψ-orthogonaux. La métrique ψ doit être considérée comme l'un des paramètres [1] dont dépendent les résultats de l'analyse : son choix est dans une large mesure arbitraire et dépend des conditions liées au domaine scientifique dans lequel s'insère le traitement statistique en cours. Les deux méthodes présentées dans ces articles ne sont que deux méthodes importantes et particulières de ce cadre. Leur importance tient au rôle que joue dans l'étude des liaisons entre les éléments de 2 ensembles :

- la corrélation linéaire dans la méthode (B) dont le calcul nécessite un protocole P application de $I \times J$ à valeurs dans R ;
- la distribution d'effectifs sur le produit cartésien $I \times J$ dans la méthode (C). Cette correspondance peut être établie pour tout protocole résultant d'une procédure de dénombrement ([22] p. 73), déterminé à partir d'un protocole de base $K \to I \times J$ même lorsque I et J sont amorphes.

INDEX DES TERMES

Deux méthodes linéaires en statistique multidimensionnelle

Plan général

1. Il en est de même de la métrique Φ dans R^n. Un bon exercice consiste à remplacer dans (II,3) l'espace (R^n, ψ) par (R^p, Φ).

II.2. *Exposé de la méthode.*

2.1. *Définitions.*

2.1.1. Correspondance.
2.1.2. Distribution des fréquences, fréquences conditionnelles.
2.1.3. Nuages. Profils et pondérations. Descriptions et descripteurs.
2.1.4. Propriétés vectorielles du nuage. Variété linéaire support du nuage.
2.1.5. Distance euclidienne sur $\mathbf{R}^p$ (resp. $\mathbf{R}^n$).

2.2. *Meilleur ajustement par un sous-espace de dimension r.*

2.2.1. Notations matricielles.
2.2.2. Recherche de la solution.
2.2.3. Propriétés des valeurs propres.
2.2.4. Axes principaux et composantes principales.

2.3. *Etude conjointe des deux nuages* $\mathcal{N}(I)$ *et* $\mathcal{N}(J)$.

2.4. *Représentations graphiques et remarques pour leur utilisation.*

2.4.1. Principes à partir desquels sont effectuées les représentations graphiques.
2.4.2. Trois propriétés utiles à l'interprétation.
2.4.3. Remarque pour l'utilisation des graphiques.
2.4.4. Part de la dispersion prise en compte par les éléments de I et J.
2.4.5. Représentation d'éléments supplémentaires et suppression d'éléments.

II.3. *Remarques et conclusion.*

Index des termes, Bibliographie.

BIBLIOGRAPHIE

Complément à la bibliographie de l'article I dans laquelle se trouvent les nos 1 à 10 absents ici.

[11a] Benzécri J.-P., *Analyse des Correspondances* Paris, Dunod, 1973.

[11b] Benzécri J.-P., « Analyse des données multidimensionnelles et classification automatique » in : *Colloque International sur la Reconnaissance de Forme*, Grenoble 1968, Edition LETI - BP 259 Grenoble, 1968, pp. 77-121.

[12] Cordier - Escofier B., Thèse de 3e cycle, *Cahiers du B.U.R.O.*, 13, ISUP, Paris, 1969, pp. 24-59.

[13] Cribier F., Deniau C., Kych A. et Lepape L., 1974, « Etude de la composition par âge de 141 villes touristiques du littoral français », *Population* n° 3.

[14a] Deniau C., Leroux B. et Oppenheim G., 1973, « Deux méthodes linéaires en statistique multidimensionnelle », *Math. Sci. hum.*, n° 44.

[14b] Deniau C. et Oppenheim G., « A propos de la factorisation des matrices : Suites complètes conjointes d'ajustement et nuages à pondérations généralisées », *Ronéo U.E.R. M.L.F.I.*, Université Paris V.1973.

[15] Faverge J.-M., 1971, « L'Analyse des Profils », *Cours polycopié de l'Université de Bruxelles.*

[16] Guttman L., Stouffer S., Suchman E., Lazarsfeld D.-P., Star S. et Clausen J., « Measurement and prediction », in : *Studies in social psychology in world war II*, Vol. IV, Princeton University Press, 1950.

[17] Hayashi C., « On the prediction of phenomena from qualitative data and the quantification of the quantitative data from the mathematical statistical point of view », *Annals of the Institute of statistical mathematics* (*Tokyo*), Vol. III, n° 2, 1952.

[18] Hayashi C., « Fundamental concept of the theory of the quantification and prediction », *Proceedings of the Institute of mathematical statistics* (*Tokyo*), Vol. III, n° 1, 1959.

[19] KEMENY J.-G. et SNELL J.-L., *Finite Markov chains*, New York, Van Nostrand, 1960.

[20] KENDALL M.-G. et STUART A., *The advanced theory of statistics*, Vol. 3., chap. 43, Londres, Griffin, 1966.

[21a] RAO C.-R., *Linear statistical inference and its application*, New York, Wiley, 1965.

[21b] RAO C.-R. et MITRA S.-K., *Generalised inverse matrices and its applications*, New York, Wiley, 1971.

[22] ROUANET H. et LEPINE D., 1972, « Notions fondamentales d'analyse des données : Protocoles », *Polycopié de l'U.E.R. M.L.F.I.* Université René Descartes (Paris V).

Nous recommandons vivement au lecteur un livre remarquable d'Algèbre Linéaire : *L'algèbre linéaire par ses applications*, Fletcher T.-J., (adapté de l'anglais par M. et V. Glaymann), Paris, CEDIC, 1972.