MATHÉMATIQUES ET SCIENCES HUMAINES

C. DENIAU
B. LEROUX
G. OPPENHEIM

Deux méthodes linéaires en statistique multidimensionnelle (1). A. - Introduction aux deux méthodes. B. - Analyse en composantes principales

Mathématiques et sciences humaines, tome 44 (1973), p. 5-34

<http://www.numdam.org/item?id=MSH_1973__44__5_0>

Math. Sci. hum. (11e année, n° 44, 1973, p. 5-34).

DEUX MÉTHODES LINÉAIRES EN STATISTIQUE MULTIDIMENSIONNELLE (1)

A. - INTRODUCTION AUX DEUX MÉTHODES
B. - ANALYSE EN COMPOSANTES PRINCIPALES

par

C. DENIAU, B. LEROUX, G. OPPENHEIM

RÉSUMÉ

Dans deux articles, dont voici le premier, sont présentés deux exemples d'analyse statistique par des méthodes factorielles. Le cadre mathématique de l'exposé est algébrique.

La présente formulation de ces problèmes s'appuie sur l'expérience d'enseignement menée à l'UER de Mathématiques, Logique Formelle et Informatique de l'Université René-Descartes, ainsi que sur une rédaction parue dans les actes du Colloque « Analyse des données en architecture et urbanisme » [5].

SUMMARY

In the two articles of which this one is the first, we present two examples of statistical analysis by factorial methods. The mathematical framework is algebraic.

The following formulation of problems rests upon the teaching experience gained at the UER de Mathématiques, Logique Formelle et Informatique de l'Université René-Descartes, and also on a publication which appeared in the proceedings of the colloquium on data analysis in architecture and urbanism [5].

0. INTRODUCTION

Pour l'étude d'un ensemble fini I d'objets à l'aide d'un ensemble fini J de I-descripteurs, on dispose d'un tableau de données X que l'on peut lire de deux manières :

Si l'on pose $|I| = n$ et $|J| = p$

$$X = \begin{bmatrix} x_1^1 & \dots & x_1^j & \dots & x_1^p \\ \cdot & & \cdot & & \cdot \\ \cdot & & \cdot & & \cdot \\ x_1^1 & \dots & x_1^j & \dots & x_1^p \\ \cdot & & \cdot & & \cdot \\ \cdot & & \cdot & & \cdot \\ x_n^1 & \dots & x_n^j & \dots & x_n^p \end{bmatrix}$$

Dans l'exemple des pages suivantes, chaque description d'un comté de l'Ohio (objet) est une suite de sept nombres ; chaque descripteur-céréale (I-descripteur) est une suite de quatre-vingt huit nombres.

Lorsque n et p sont grands, ce tableau est difficilement déchiffrable.

L'objet des méthodes présentées ici est de fournir à partir du tableau d'étude un outil de lecture des liens entre les éléments de I, entre les éléments de J et enfin entre ceux de I et de J.

Deux points de vue sur ces méthodes sont classiques :

i) Pour illustrer le premier, nous empruntons à un texte déjà ancien de C. R. Rao [8] la formulation suivante : « quand on dispose d'un grand nombre de mesures (*measurements*), il est naturel de se demander s'il est possible de les remplacer par un nombre plus petit d'entre elles (ou par un nombre plus petit de fonctions de ces mesures), sans que trop d'informations soit perdue... ».

Pour cela, on recherche un ***bon résumé*** du tableau X. Il s'agit, par des opérations algébriques simples (ajustement linéaire), de déterminer à partir des I-descripteurs initiaux, un *petit nombre* q ($q \ll p$) de nouveaux I-descripteurs ou *composantes principales* (fonctions linéaires des précédents).

Les I-descripteurs doivent :

a) être non corrélés deux à deux ;

b) engendrer un sous-espace de dimension q prenant en compte une part *suffisante* de la *dispersion* du nuage des objets, part *maximum* en q I-descripteurs. Si cette part n'est pas jugée suffisante on prend plus de q I-descripteurs.

$$X \mapsto Y = \begin{bmatrix} y_1^1 & \dots & y_1^q \\ & & \\ y_n^1 & \dots & y_n^q \end{bmatrix} \quad q \ll p$$

ii) Le second point de vue consiste, s étant le rang de la matrice ($s \leqslant p$), à déterminer s I-descripteurs (ou composantes principales) nouveaux, fournissant des descriptions meilleures que les descriptions initiales dans le sens suivant :

a) ils sont non corrélés deux à deux ;

b) ils sont de dispersion décroissante ; chacun d'entre eux prenant en compte une part maximum de la dispersion.

Pratiquement on ne fait rien de plus que d'effectuer un changement de repère orthonormal dans le sous-espace de $\mathbb{R}^p$ dans lequel sont représentés les objets.

Remarquons qu'en prenant les q premiers nouveaux I-descripteurs on obtient le meilleur résumé en q descripteurs au sens du premier point de vue.

Le choix entre les deux points de vue s'effectue en fonction :

— des questions que se pose le chercheur ;

— de la taille du tableau X ;

— des moyens de calcul.

Lorsque le nombre des I-descripteurs n'est pas trop grand, on a tout intérêt à adopter le deuxième point de vue.

Dans le cas de très grands tableaux de données, n *et* p grands, il est alors suffisant de déterminer quelques nouveaux I-descripteurs. Il est alors nécessaire de posséder des outils permettant d'apprécier la qualité du résumé obtenu ainsi. Deux notions sont introduites ici :

a) la qualité de l'ajustement (ou *qualité du résumé*) ;

b) la qualité de la représentation des objets.

Vues dans leur ensemble ces méthodes permettent en outre :

— de hiérarchiser le rôle de chacun des nouveaux I-descripteurs selon l'importance de leur pouvoir discriminant (qui est ici par définition leur variance). Ainsi la première composante principale est celle qui permet la meilleure discrimination des objets étudiés ;

— de constituer des *classes* d'objets ou de I-descripteurs proches ;

— d'apprécier l'hétérogénéité ou l'homogénéité de l'ensemble des objets étudiés et des classes constituées ;

— de déterminer la ou les propriétés (en terme de descripteurs) communes à des éléments constituant une classe.

Des *représentations graphiques* permettent de *visualiser* partiellement les résultats que l'on vient de décrire.

La méthode de détermination des nouveaux I-descripteurs est basée sur la proposition suivante justifiée au paragraphe 2.3 : la recherche des I-descripteurs non corrélés de variance décroissante se ramène à la détermination de nouvelles descriptions conservant le mieux possible en moyenne les distances calculées entre les descriptions initiales (la somme des carrés des distances entre tous les couples de lignes de Y, voisine de la même somme calculée sur X).

La résolution numérique de ce problème (décomposition spectrale du tableau X) et celle de problèmes connexes, surtout dans les cas intéressants dans lesquels n et p sont grands, est récente.

En effet, si les bases *mathématiques* de ces techniques sont déjà anciennes, les calculs numériques (calculs des valeurs propres de grandes matrices) nécessaires à leur mise en œuvre et à leur développement n'ont été possibles qu'avec la naissance des gros calculateurs et les progrès de l'analyse numérique.

1. EXEMPLE D'ANALYSE EN COMPOSANTES PRINCIPALES

a) Les données

Elles constituent une petite part des données utilisées par J. C. Weaver [9] et [10] pour une étude régionale des différentes cultures dans les 1 081 comtés des états du « Middle-West » (USA). L'auteur détermine, à l'aide de techniques simples, des régions se distinguant les unes des autres par des types différents de combinaisons de cultures ; il compare deux classifications obtenues sur des données relatives aux années 1939 et 1949. Seules les données de 1949 pour les 88 comtés de l'état de l'Ohio ont été retenues, et parmi les cultures ne figurent que les sept céréales et plantes fourragères suivantes :

maïs, petites graines, blé, avoine, soja, foin, orge.

(Pour plus de détails concernant l'ensemble des cultures et des exploitations agricoles étudiées, se reporter à [9] et [10] où l'on trouvera aussi des cartes résumant les résultats de l'étude.)

Chacun des 88 comtés de l'état de l'Ohio est décrit par le pourcentage de surface agricole exploitée en chacune des sept céréales et plantes fourragères, rapporté à la surface totale exploitée. (On tient compte, dans le calcul de ce total, d'autres cultures telles que le lin, le coton, etc.)

On peut alors dire qu'à tout comté on associe une suite de 7 nombres réels : sa *description.* De même à chaque céréale ou plante fourragère on associe une suite de 88 nombres réels (pourcentage de surface exploitée en cette céréale dans chacun des comtés) : on l'appellera un I-descripteur (Annexe II).

Le Tableau 1 fournit quelques exemples de descriptions de comtés ; pour obtenir la totalité des données que nous avons utilisées, en particulier les I-descripteurs se reporter à [7].

b) L'analyse

Elle est effectuée sur un tableau de données, *transformé* du tableau initial, de la manière suivante :

i) les I-descripteurs (soit ici les colonnes du tableau initial) sont ***centrés*** ;

ii) les mêmes I-descripteurs sont ensuite ***réduits*** (important) ;

iii) toutes les descriptions sont affectées de la *même* masse.

Index des tableaux de l'exemple

1. Description de quelques comtés.
2. Moyenne et écart-type des I-descripteurs.
3. Matrice des corrélations.

4. Valeurs propres de la matrice des corrélations.
5. Vecteurs propres de la matrice des corrélations.
6. Corrélations entre les I-descripteurs et les composantes principales.
7. Projection de quelques comtés sur les quatre premiers axes factoriels.
8. Qualité de la représentation de quelques comtés.

Tableau 1. *Descriptions de quelques comtés*

Comté	Maïs	Petites graines	Blé	Avoine	Orge	Soja	Foin
Adams	42.41	0.21	22.47	1.07	0.37	0.62	27.80
Allen	34.43	0.13	23.76	18.35	0.11	12.18	15.31
Athens	26.61	0.18	8.89	3.42	0.05	0.71	53.91
Auglaze	34.66	0.06	22.02	18.24	0.10	9.80	11.97
Belmont	18.60	0.11	10.87	9.16	0.31	0.20	50.82
Clinton	48.45	0.24	29.50	3.10	0.25	2.72	9.85
Fayette	41.63	0.14	26.65	5.75	0.27	7.59	11.79
Gallia	31.38	0.83	13.07	2.03	0.60	0.71	44.07
Geauga	23.04	0.21	12.68	17.44	0.11	0.41	37.80
Greene	49.23	0.23	29.30	4.63	0.31	1.97	15.78
Hancock	36.13	0.12	24.64	16.56	0.13	13.91	16.46
Hardin	33.07	0.15	16.63	14.61	0.13	13.18	14.18
Putnam	30.97	0.13	24.16	15.28	0.13	14.10	12.61
Sandusky	30.37	0.09	20.96	14.24	0.16	15.20	14.29
Seneca	31.31	0.04	25.51	14.38	0.12	10.76	15.43
Shelby	36.53	0.05	22.33	15.42	0.03	9.88	14.96
Warren	43.23	0.09	24.97	3.20	0.24	4.68	18.72

Tableau 2. *I-descripteurs*

	Maïs	Petites graines	Blé	Avoine	Orge	Soja	Foin
Moyenne	31.78	0.14	20.16	10.39	0.16	6.08	25.11
Écart-type	8.67	0.11	5.68	6.46	0.20	6.78	13.78

Matrice des corrélations

Maïs	1.000						
Petites graines	— 0.033	1.000					
Blé	0.463	— 0.124	1.000				
Avoine	— 0.328	— 0.081	0.075	1.000			
Orge	0.148	0.249	— 0.044	— 0.255	1.000		
Soja	0.010	0.005	0.151	0.396	— 0.201	1.000	
Foin	— 0.436	0.104	— 0.648	— 0.330	0.109	— 0.601	1.000

Tableau 3. *Valeurs propres de la matrice des corrélations*

2.34	Qualité de l'ajustement en pourcentage	33.48
1.69	Qualité de l'ajustement en pourcentage	24.15
1.13	Qualité de l'ajustement en pourcentage	16.15
0.694	Qualité de l'ajustement en pourcentage	9.93
0.613	Qualité de l'ajustement en pourcentage	8.77
0.370	Qualité de l'ajustement en pourcentage	5.29
0.156	Qualité de l'ajustement en pourcentage	2.24

Tableau 4. *Vecteurs propres de la matrice des corrélations*

	Maïs	Petites graines	Blé	Avoine	Orge	Soja	Foin
u_1	0.28	— 0.15	0.47	0.28	— 0.18	0.44	— 0.61
u_2	— 0.59	— 0.11	— 0.30	0.53	— 0.41	0.28	0.08
u_3	— 0.10	0.78	— 0.11	0.18	0.44	0.34	— 0.13
u_4	— 0.20	— 0.47	0.09	0.39	0.074	— 0.14	— 0.05
u_5	— 0.15	0.393	0.545	0.355	— 0.223	— 0.588	0.06
u_6	0.645	0.0331	— 0.49	0.522	— 0.065	— 0.258	— 0.011
u_7	0.284	— 0.0306	0.342	0.195	0.365	0.408	0.771

Tableau 5. *Corrélations entre les I-descripteurs et les composantes principales*

Composantes principales	1°	2°	3°	4°
Maïs	0.428	— 0.770	— 0.104	— 0.167
Petites graines	— 0.224	— 0.147	0.834	— 0.391
Blé	0.726	— 0.400	— 0.121	0.079
Avoine	0.428	0.694	0.190	0.325
Orge	— 0.275	0.538	0.473	0.621
Soja	0.669	0.362	0.361	— 0.117
Foin	— 0.934	0.107	— 0.136	— 0.038

Tableau 6. *Projection de quelques comtés sur les quatre premiers axes factoriels* (Avec la totalité des comtés, les quatre colonnes fourniraient les quatre premières composantes principales)

	Axe 1	Axe 2	Axe 3	Axe 4
Adams	— 0.617	— 2.338	0.207	— 0.186
Allen	1.629	0.603	— 0.292	0.256
Athens	— 2.977	0.532	— 0.435	— 1.048
Auglaze	1.581	0.649	— 0.304	0.547
Belmont	— 2.856	0.951	— 0.160	0.813
Clinton	1.246	— 2.772	0.279	— 0.646
Fayette	1.253	— 1.648	0.027	0.030
Gallia	— 3.485	— 2.020	5.468	— 1.959
Geauga	— 1.574	1.472	0.404	— 0.100
Greene	0.973	— 2.796	0.283	— 0.313
Hamilton	— 1.447	— 2.736	0.643	3.275
Hancock	1.739	0.338	0.253	0.201
Putnam	1.642	0.587	0.404	0.200
Sandusky	1.335	0.753	0.260	0.356
Seneka	1.518	0.409	— 0.530	0.580
Shelby	1.496	0.445	— 0.662	0.110
Warren	0.655	— 1.843	— 0.672	— 0.043

Tableau 8. *Qualité de la représentation de quelques comtés*

	Axe 1	Axe 2	Axes 1-2	Axes 1-2-3
Adams	0.064	0.922	0.986	0.993
Allen	0.771	0.105	0.877	0.902
Athens	0.817	0.026	0.843	0.861
Auglaze	0.688	0.116	0.804	0.830
Belmont	0.824	0.091	0.915	0.918
Clinton	0.152	0.754	0.907	0.914
Fayette	0.357	0.618	0.976	0.976
Franklin	0.028	0.114	0.143	0.833
Gallia	0.235	0.079	0.314	0.893
Geauga	0.414	0.363	0.778	0.805
Greene	0.099	0.822	0.922	0.930
Putnam	0.800	0.102	0.902	0.951
Sandusky	0.569	0.181	0.751	0.772
Seneca	0.707	0.051	0.758	0.844
Shelby	0.713	0.063	0.776	0.916
Warren	0.097	0.771	0.869	0.972

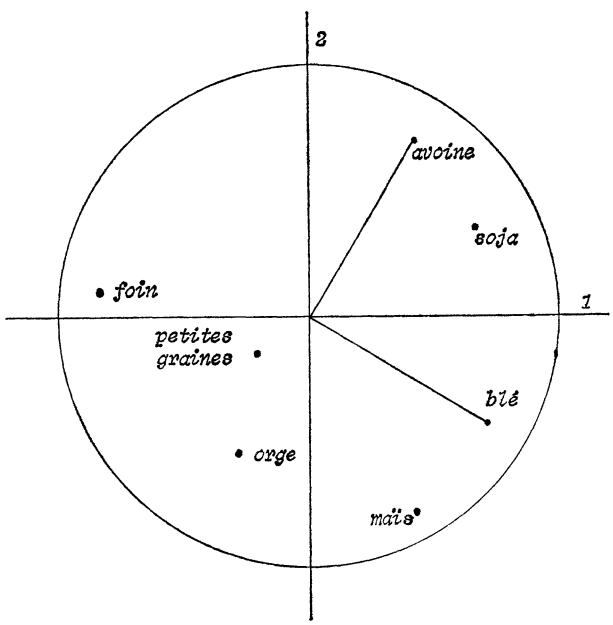

Figure 1. *Représentation dans le plan des deux premières composantes principales* (*premier axe* 33,48 % ; *deuxième axe* 24,15 %)

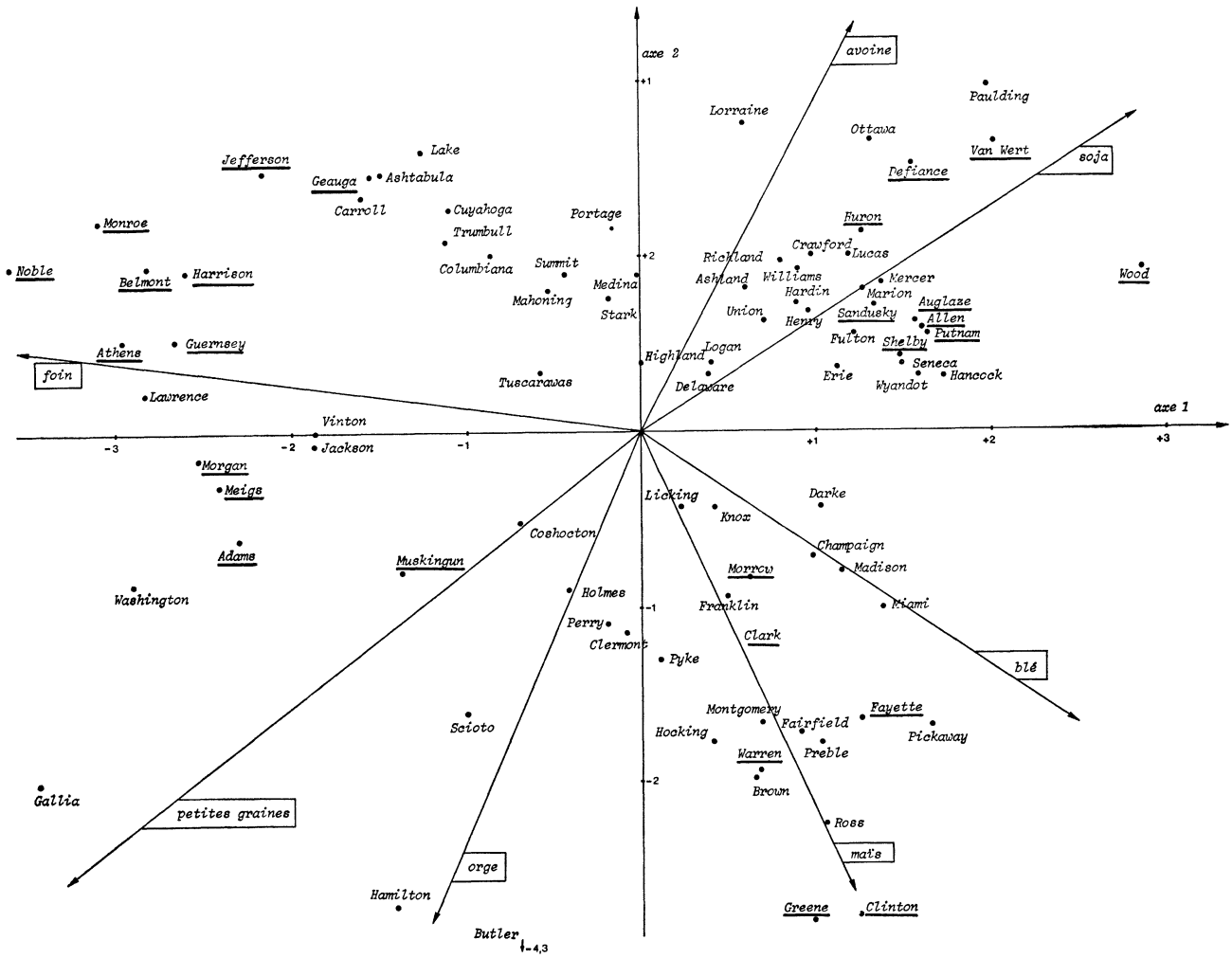

Figure 2

c) *Quelques interprétations des résultats fournis par l'analyse*

Les remarques présentées ci-dessus sont des remarques très générales illustrant le présent exemple.

Les données analysées ont été préalablement *centrées et réduites* en colonne. Le Tableau 1 donne les moyennes et les écarts-types des I-descripteurs, avant *centrage et réduction.* L'écart-type « petites graines » est de l'ordre de 0.11, alors que celui du foin est de 13.78, soit plus de *cent fois* supérieur.

1) *Les composantes principales et la Figure* 1

Le *Tableau* 6 fournit les corrélations entre les I-descripteurs initiaux et les quatre premières composantes principales.

— le foin [1] est fortement corrélé négativement avec la première composante principale ;

— le blé, le maïs et l'avoine sont bien représentés par (corrélés avec) le plan vectoriel des deux premières composantes principales ;

— les petites graines ne sont corrélées avec aucune des deux premières composantes principales.

La Figure 1 permet de visualiser partiellement le Tableau 5 : chaque point du plan est repéré par ses corrélations avec la première et la deuxième composante principale. Un simple coup d'œil sur ce graphique donne une bonne idée des deux premières colonnes du Tableau 5. Le cercle dessiné a pour rayon un.

Blé et avoine sont bien représentés dans ce plan, l'angle de ces deux vecteurs est voisin de $\pi/2$ et leur corrélation 0.075 se lit sur le Tableau 3.

On aurait pu construire d'autres graphiques pour d'autres composantes principales (ex. : 1re et 3e composantes principales à l'aide duquel pourrait se lire la forte corrélation entre « petites graines » et cette 3e composante principale).

2) *Le nuage des comtés et la Figure* 2

La part et la variance du nuage des comtés prise en compte par le plan des deux premiers axes factoriels est supérieure à 57 %. Cela fournit un bon indice de l'intensité de ces deux directions de dispersion du nuage.

La Figure 2 donne une visualisation des colonnes 1 et 2 des Tableaux 4, 6, 7 ; elle fournit les projections des différents comtés sur le plan des deux premiers axes factoriels (les comtés soulignés sont bien représentés, ainsi que les projections des directions (orientées) des vecteurs de la base initiale (blé, avoine, ...).

Le premier axe factoriel sépare les comtés, dont la part en foin est importante, des autres comtés. Une première coordonnée fortement négative caractérise les comtés dont la part en foin est importante, celle en blé et soja faible. Les proximités entre *Athens, Belmont, Clinton* sont lues directement sur ce graphique (ces points sont bien représentés), ils se situent tous à l'est de l'Ohio.

Le deuxième axe factoriel sépare les comtés de la façon suivante :

— ceux dont la part en maïs et blé (céréales principales) est importante (coordonnées, relativement à ce deuxième axe, négatives) ;

— ceux dont la part en avoine (céréale fourragère) est importante (coordonnées positives, relativement à ce deuxième axe).

Le troisième axe factoriel sépare les comtés ayant une part importante en petites graines et orge ou en soja, des autres (si l'on n'avait pas centré et réduit les colonnes du tableau des données initiales, petites graines et orge auraient très probablement eu un pouvoir discriminant négligeable).

1. Lorsque l'on nomme une des plantes on sous-entend « la part de surface agricole exploitée par cette plante ».

2. BASES MATHÉMATIQUES DES MÉTHODES FACTORIELLES

Ce chapitre a pour but de préciser quelques notions utiles dans la présentation des méthodes factorielles. Les mots clés sont :

Nuage. Projection orthogonale. Dispersion. Ajustement linéaire. Qualité de l'ajustement. Qualité de la représentation.

Dans tout ce paragraphe $E = \mathbb{R}^p$ sera un espace euclidien. On notera $(u|x)$ le produit scalaire, des deux vecteurs u et x, sans plus de précision sur ce produit scalaire, et Ψ la matrice associée à ce produit scalaire relativement à la base canonique de E.

2.1. *Nuage de points et nuage projeté*

Définition (2.1.1) *: Nuage de points*

On appelle nuage de points de E munis de masses ponctuelles un ensemble fini de couples $\mathscr{N} = (x_i, m_i)_{i \in I}$ où I est un ensemble fini, et quel que soit $i \in I$, x_i est un vecteur de E et m_i un élément de $\mathbb{R}$ non négatif (une masse). On suppose évidemment que les m_i sont non tous nuls.

Remarque : On dira très souvent : x_i élément du nuage $\mathscr{N}$, même si cela constitue un abus de langage et on notera parfois $\mathscr{N} \subset E$.

Remarque (2.1.2) *:* Le plus souvent les masses sont telles que $\sum_{i=1}^{n} m_i = 1$; enfin il est fréquent que toutes ces masses soient égales : $m_i = \frac{1}{n}$.

Point moyen du nuage : $M = M(\mathscr{N}) = \dfrac{\sum_{i=1}^{n} m_i x_i}{\sum_{i=1}^{n} m_i}$.

Un nuage $\mathscr{N}$ est centré si $M = 0$.

Définition (2.1.2) : *Nuage projeté et nuage résiduel*

Soit V un sous-espace vectoriel de E, $\mathscr{N}$ un nuage de E et p_V l'application projection orthogonale sur V. On appellera *nuage projeté* sur V le nuage $\mathscr{N}_V = p_V(\mathscr{N}) = (p_V(x_i), m_i)_{i \in I}$. On appellera *nuage résiduel* associé à $\mathscr{N}$ et V le nuage $\mathscr{N}_{V^\perp}$ où $V^\perp$ est le supplémentaire orthogonal de V dans E.

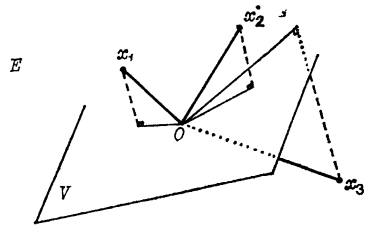

Figure 3. *Nuage projeté sur un sous-espace vectoriel V*

Cas particulier : Nuage projeté sur une droite vectorielle.

Soit $V = [u]$, la droite vectorielle engendrée par un vecteur unitaire $u \in E$, donnons une expression simple et utile de la projection orthogonale :

$$p_{[u]}(x) = (u|x)\,u$$

et

$$p_{[u]}(\mathscr{N}) = ((u|x_i)\,u,\ m_i)_{\ i \in I}$$

expression dans laquelle, quel que soit $i \in I$, $(u|x_i)$ est la coordonnée de $p_{[u]}(x_i)$ par rapport au vecteur unitaire u.

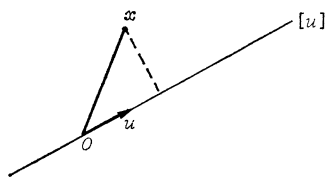

Figure 4. *Cas particulier à* $V = [u]$

2.2. *Dispersion du nuage et dispersion du nuage projeté*

Soit $v = ((v_i, m_i))_{\ i \in]n]}$ une suite v_i de nombres réels affectés d'une masse m_i avec $\sum_{i=1}^{n} m_i = 1$.

Un indice usuel de dispersion est fourni par la variance de cette suite :

$$Var\ v = \sum_{i=1}^{n} m_i\ [v_i - M(v)]^2 \qquad \text{avec } M(v) = \sum_{i=1}^{n} m_i\ v_i.$$

On reconnaît la moyenne des carrés des distances des v_i à leur moyenne pondérée M(v). Pour un nuage de E cette notion se généralise de la façon suivante :

Définition (2.2.1)

On appelle dispersion du nuage $\mathscr{N} = ((x_i, m_i))_{\ i \in I}$ et l'on note *Disp* $\mathscr{N}$ l'expression

$$Disp\ \mathscr{N} = \sum_{i=1}^{n} m_i\ \|x_i - M\|^2.$$

Cette quantité est la moyenne pondérée des carrés des distances au point moyen du nuage.

Autre formulation de la dispersion d'un nuage

Comme il est possible de le faire pour la variance d'une suite de nombres réels, exprimons la dispersion en ne faisant intervenir que les *distances entre les couples des vecteurs du nuage*. Dans cette expression le point moyen ne joue pas de rôle particulier.

Soit $\mathscr{G} = ((\mathbf{x}_i, m_i))_{i \in 1}$ on a :

$$\sum_{i=1}^{n} m_i \, \|\mathbf{x}_i - M\|^2 = \frac{1}{2 \left(\sum_{i=1}^{n} m_i\right)} \sum_{i=1}^{n} \left[\sum_{j=1}^{n} m_i m_j \, \|\mathbf{x}_i - \mathbf{x}_j\|^2 \right]$$

Nous laissons au lecteur le soin de montrer cette propriété. Dans le cas particulier où $\sum_{i=1}^{n} m_i = 1$, le coefficient $\dfrac{1}{2 \sum_{i=1}^{n} m_i}$ est égal à $\dfrac{1}{2}$.

Remarque (2.2.1)

De même que la variance d'une suite finie de nombres réels (nuage porté par une droite vectorielle) ne caractérise pas la suite, de même la dispersion ne fournit pas la « forme » du nuage. Deux nuages très différents peuvent avoir même dispersion. Soient les 2 nuages $\mathscr{G}_1$ et $\mathscr{G}_2$:

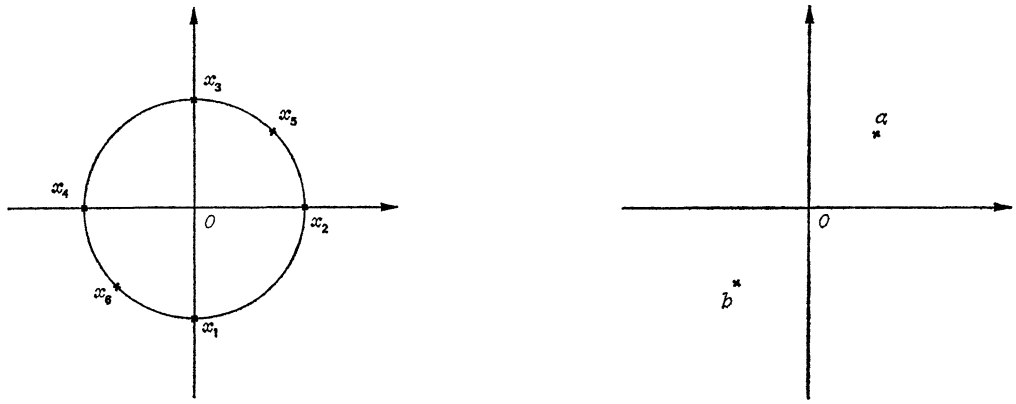

Figure 5

5.1. *Nuage* $\mathscr{G}_1$ sur la sphère unité de $\mathbb{R}^2$ 5.2. *Nuage* $\mathscr{G}_2$ *défini par* $\begin{cases} \forall\ i \in]3] : \mathbf{x}_i = a \\ \forall\ i \in [4,6] : \mathbf{x}_i = b \\ d(0,a) = d(0,b) = 1 \end{cases}$

Chacun des six vecteurs est muni de la masse $m = 1/6$.

On a : *Disp* $\mathscr{G}_1 = 1$ et *Disp* $\mathscr{G}_2 = 1$.

Il peut donc être intéressant pour étudier la forme du nuage de considérer la dispersion dans certaines directions. La dispersion du nuage projeté sur une droite vectorielle est définie par :

Dispersion du nuage projeté et dispersion résiduelle. Étant donné un nuage $\mathscr{N}$ de E et V un sous-espace vectoriel, on note *Disp* $\mathscr{N}_V$ la dispersion du nuage projeté $\mathscr{N}$. La ***dispersion résiduelle*** est, par définition, la dispersion du nuage résiduel $\mathscr{N}_V{}^+$ associé.

Cas particulier : V est une droite vectorielle.

Si $V = [u]$ où u est un vecteur unitaire :

$$\mathscr{N}_{[u]} = ((u|x_i)\,u,\ m_i)_{\ i \in I}$$

la dispersion du nuage projeté est :

$$\begin{aligned} \textit{Disp}\ \mathscr{N}_{[u]} &= \sum_{i=1}^{n} m_i\ \|(u|x_i)\,u - (u|M)\,u\|^2 \\ &= \sum_{i=1}^{n} m_i\ \|(u|x_i - M)\,u\|^2 \\ &= \sum_{i=1}^{n} m_i\ (u|x_i - M)^2 \\ &= S(u) \end{aligned}$$

Donc à la dispersion du nuage projeté sur [u] peut être associée la restriction à la sphère-unité de E d'une forme quadratique définie sur E. On l'appelle *forme quadratique de dispersion* du nuage $\mathscr{N}$. On notera s la forme bilinéaire symétrique associée à S, on l'appellera forme bilinéaire de dispersion. Enfin on note Σ la matrice de s relativement à la base canonique de $\mathbb{R}^p$.

Décomposition de la dispersion sur deux sous-espaces orthogonaux

Soient V et V^+ un couple de sous-espaces vectoriels supplémentaires orthogonaux de E.

$$\textit{Disp}\ \mathscr{N} = \textit{Disp}\ \mathscr{N}_V + \textit{Disp}\ \mathscr{N}_V{}^+$$

Soit $\{w_i\ /\ i \in]n]\}$ une base orthonormée de E :

$$\textit{Disp}\ \mathscr{N} = \sum_{i=1}^{n} \textit{Disp}\ \mathscr{N}_{[w_i]}$$

2.3. *Ajustement linéaire*

Définition (2.3.1). Étant donné $\mathscr{N}$ un nuage de E et *Disp* $\mathscr{N}$ sa dispersion, on appelle *ajustement linéaire* de la dispersion du nuage par un sous-espace vectoriel V de E le couple ($\mathscr{N}_V$, *Disp* $\mathscr{N}_V$) constitué par le nuage projeté orthogonalement sur V et par la dispersion de ce nuage.

Lorsqu'il n'y a pas d'ambiguïté nous parlons simplement d'ajustement du nuage.

Définition (2.3.2) : *Qualité de l'ajustement.* On appelle qualité de l'ajustement du nuage $\mathscr{N}$ par un sous-espace vectoriel V le quotient de la dispersion du nuage projeté sur V par la dispersion du nuage $\mathscr{N}$.

On note $Q_V = \dfrac{Disp\ \mathscr{N}_V}{Disp\ \mathscr{N}}$, on a $0 \leqslant Q_V \leqslant 1$;

Remarque (2.3.1) : Dans l'exemple de la remarque (2.2.1) on constate que $\mathscr{N}_2$ est porté par une droite V de $\mathbb{R}^2$: la dispersion de $\mathscr{N}_2$ et la dispersion du nuage projeté sur cette droite sont égales, alors $Q_V = 1$.

On aurait pu construire un nuage $\mathscr{N}_3$ dont les éléments sans être sur la droite en seraient assez peu écartés. Alors $Q_V \simeq 1$, et la connaissance de la droite et de Q_V nous fournirait un bon indice de la forme du nuage.

Définition (2.3.3) : Étant donnés deux sous-espaces V_1 et V_2 et un nuage $\mathscr{N}$ de E, on dit que l'ajustement de $\mathscr{N}$ par V_1 est meilleur que celui par V_2 si :

$$Q_{V_1} \geqslant Q_{V_2}.$$

Remarque (2.3.2) : $\mathscr{N}$ étant donné, le dénominateur de Q_V est une constante, aussi on peut dire que l'ajustement de $\mathscr{N}$ par V_1 est meilleur que celui par V_2 si : $Disp\ \mathscr{N}_{V_1} \geqslant Disp\ \mathscr{N}_{V_2}$.

Remarque (2.3.3) : $Q_V = 0$ si et seulement si tout vecteur x_i du nuage est dans $V^{\perp}$.

$Q_V = 1$ si et seulement si tout vecteur x_i du nuage est dans V.

2.4. *Recherche du meilleur ajustement*

Étant donné un nuage $\mathscr{N}$ de E, quel est parmi tous les sous-espaces de dimension r de E, celui ou ceux qui fournissent le meilleur ajustement de $\mathscr{N}$?

a) Décomposition de la dispersion du nuage projeté sur V

Théorème (2.4.1) :

— soient E un espace vectoriel de dimension supérieure à 1 et $\mathscr{N}$ un nuage de E ;

— soit $\{u_i \ / \ i \in]r]\}$ un ensemble de droites vectorielles de E telles que :

i) *La*[1] droite $[u_1]$ réalise le meilleur ajustement de dimension 1 de $\mathscr{N}$.

---

1. Nous supposons ici l'unicité de $[u_1]$, $[u_2]$, ..., $[u_r]$ et de V. Cette hypothèse n'est pas nécessaire pour démontrer le théorème, et il suffit d'un petit effort pour obtenir l'énoncé du théorème et sa démonstration dans le cas général.

Nous ne le faisons pas pour deux raisons :

a) Les principes des deux démonstrations sont les mêmes, mais la démonstration du cas particulier est plus courte et ne nécessite aucune notation compliquée.

b) Dans la pratique, l'unicité des $[u_i]$ est la règle générale.

Cette hypothèse d'unicité entraîne évidemment : pour tout i, j élément de $]r]$, $i < j$ implique $S(u_i) > S(u_j)$.

ii) Quel que soit $k \in [2, r]$, parmi toutes les droites orthogonales à $\bigoplus_{i=1}^{k-1} [u_i]$ *la* droite $[u_k]$ réalise le meilleur ajustement de dimension un de $\mathscr{C}$;

— soit V *le* sous-espace vectoriel de dimension $r \in]n]$ de E qui réalise le meilleur ajustement de dimension r de $\mathscr{C}$

alors $\boxed{V = \bigoplus_{i=1}^{r} [u_i]}$

(preuve en annexe I)

Corollaire (2.4.1)

Sous les mêmes hypothèses : $Disp\ \mathscr{C}_V = \sum_{i=1}^{r} S(u_i)$.

En effet : $Disp\ \mathscr{C}_V = \sum_{i=1}^{r} m_i \|p_V(x_i - M)\|^2$,

Or si $\{u_i \ / \ i \in]r]\}$ est une base orthonormée de V,

$$\|p_V(x_i - M)\|^2 = \sum_{i=1}^{r} \|p_{[v_i]}(x_i - M)\|^2.$$

donc : $Disp\ \mathscr{C}_V = \sum_{i=1}^{r} S(u_i)$.

b) *Résolution du problème*

• Comment déterminer $[u_1]$?

La droite $[u_1]$ réalise le meilleur ajustement par un sous-espace de dimension 1.

$$S(u_1) = \sum_{i=1}^{n} m_i (u_1 \mid x_i - M)^2 = Sup_{\|u\|=1} \left(\sum_{i=1}^{n} m_i (u \mid x_i - M)^2 \right).$$

On est ramené à résoudre un problème connu : déterminer le maximum d'une forme quadratique sur la sphère unité de E :

1) u_1 [1] est un vecteur propre unitaire de la forme quadratique associée à λ_1 plus grande valeur propre de la matrice $\Psi^{-1}\,\Sigma$.

2) $S(u_1) = \lambda_1$.

---

1. Nous confondons u_1 et la matrice colonne de u_1 dans la base canonique de $\mathbb{R}^p$.

• Comment déterminer $[u_2]$?

La droite $[u_2]$ est parmi les droites orthogonales à $[u_1]$ celle qui réalise le meilleur ajustement :

$$S(u_2) = \underset{\substack{\|u\| = 1 \\ u \perp u_1}}{Sup} \left(\sum_{i=1}^{n} m_i \, (u \mid x_i - M)^2 \right).$$

On sait que :

1) u_2 est un vecteur propre unitaire de la matrice $\Psi^{-1} \Sigma$ associé à λ_2, plus grande valeur propre de $\Psi^{-1} \Sigma$ inférieure à λ_1.

2) $S(u_2) = \lambda_2$.

Et ainsi de suite. Pour une démonstration se reporter à [4] ou [8].

c) Retour sur la qualité de l'ajustement

Soient $\lambda_1, \lambda_2, ..., \lambda_p$, les valeurs propres de $\Psi^{-1} \Sigma$.

Soit V un sous-espace de dimension r sur lequel le nuage se projette avec une dispersion maximum ;

$$Q_V = \frac{\text{Dispersion du nuage projeté sur V}}{\text{Dispersion du nuage}}$$

On a

$$Q_V = \frac{\sum_{i=1}^{r} \lambda_i}{\sum_{i=1}^{p} \lambda_i} = \frac{\sum_{i=1}^{r} \lambda_i}{Tr(S)} \quad .$$

Expression et interprétation de la matrice Σ

Par rapport aux bases canoniques de $\mathbb{R}^p$ et $\mathbb{R}^n$, notons :

X la matrice $n \times p$ des données associée au nuage $\mathscr{N}$.

$\theta = (m_i \, \delta_{ii'})_{1 \leqslant i,i' \leqslant n}$ une matrice diagonale $n \times n$ dont les éléments sont les masses (déf. 2.1.1.).

M la matrice colonne associée au point moyen M du nuage $\mathscr{N}$ dans $\mathbb{R}^p$

$$\delta_n = (1, 1, ..., 1) \in \mathbb{R}^n$$

On montre alors facilement que la matrice Σ associée à la forme bilinéaire symétrique de dispersion est :

$$\Sigma = {}^t\tilde{Y} \, \theta \, \tilde{Y}$$

où $Y = X \Psi$ et $\tilde{Y} = Y - \delta_n \, {}^tM \, \Psi$.

Si quel que soit $i \in]n]$, $m_i > 0$, on peut interpréter θ comme la matrice associée à un produit scalaire de $\mathbb{R}^n$, noté aussi θ.

Alors le terme général $s_{jj'}$ de la matrice Σ est

$$s_{jj'} = \theta(y^j, y^{j'}) = \mathrm{cov}_\theta(y^j, y^{j'})$$

où quel que soit $j \in]p]$, y^j et $y^{j'}$ sont les vecteurs de $\mathbb{R}^n$ dont les coordonnées par rapport à la base canonique sont respectivement les éléments de la $j^{\text{ème}}$ et de la $j'^{\text{ème}}$ colonnes de Y.

2.5. *Qualité de la représentation*

Dans l'exemple de la remarque (2.2.1) le nuage $\mathscr{N}_2$ de $\mathbb{R}^2$ est situé sur une droite. Si l'on ajuste cette droite au nuage on constate que les distances des points du nuage entre eux et des points du nuage projeté entre eux sont les mêmes.

Si le nuage avait été approximativement sur cette droite les distances auraient été peu différentes.

Par contre, quel que soit l'ajustement par une droite du nuage $\mathscr{N}_1$, certaines distances entre couples de points et couples de points projetés vont être différentes.

Remarque (2.5.1) : L'ajustement est effectué à partir de la dispersion *Disp* $\mathscr{N}$ qui est comme nous l'avons vu en (2.2) un indice global. Même si l'ajustement réalisé est le meilleur, il n'est pas optimal pour chacun des vecteurs. S'il paraît vraisemblable qu'un grand nombre de vecteurs sont bien représentés, il est intéressant d'étudier individuellement la position [1] de x_i par rapport au sous-espace V. Cette étude est particulièrement importante dans le cas d'un résumé. En effet, même si la dispersion résiduelle est faible, cela ne prouve pas que le résumé soit aussi bon pour chacun des points. Des points sont alors mal représentés par le résumé. Il est nécessaire de diagnostiquer la cause de cette hétérogénéité.

Définition (2.5.1) : Pour tout vecteur x de E, $(x \neq 0)$, on appelle qualité de la représentation du vecteur x par le sous-espace V de E le rapport :

$$q_V(x) = \frac{\|p_V(x)\|^2}{\|x\|^2} = cos^2\theta.$$

- $q_V(x)$ est le coefficient de corrélation multiple de x et de V.
- $0 \leqslant q_V(x) \leqslant 1$.
- $q_V(x) = 1 \Leftrightarrow x \in V$ et $q_V(x) = 0 \Leftrightarrow x \in V^\perp$.
- Plus petit est l'angle entre x et V, plus grande est la qualité de la représentation de x par V.
- Si $V = V_1 \oplus V_2$ et $V_1 \perp V_2$: $q_V(x) = q_{V_1}(x) + q_{V_2}(x)$.

1. Il peut être intéressant de comparer l'indice défini en (2.5.1) à celui défini par Rao dans [8] page 334.

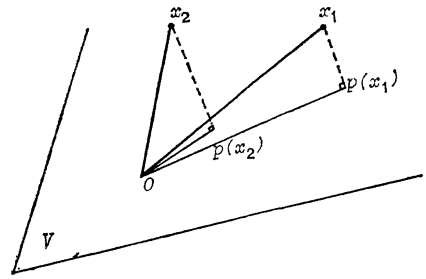

Figure 6. *Qualité de la représentation*

Pour un tour plus complet de la question nous renvoyons le lecteur à [6].

3. ANALYSE EN COMPOSANTES PRINCIPALES

3.1. *Introduction*

3.1. *a*) Revenons à l'étude d'un ensemble I d'objets à l'aide d'un ensemble $(\delta_j)_{j \in J}$ de descripteurs.

Posons : $|I| = n$, $|J| = p$

$$\mathbf{X} = \begin{bmatrix} x_1^1 & \dots & x_1^j & \dots & x_1^p \\ \vdots & & \vdots & & \vdots \\ x_i^1 & \dots & x_i^j & \dots & x_i^p \\ \vdots & & \vdots & & \vdots \\ x_n^1 & \dots & x_n^j & \dots & x_n^p \end{bmatrix} \begin{matrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_i \\ \vdots \\ \mathbf{x}_n \end{matrix}$$

$$\mathbf{x}^1 \ \dots \ \mathbf{x}^j \ \dots \ \mathbf{x}_p \quad ;$$

à l'objet $i \in I$ est associée sa description : $(x_i^1, \dots, x_i^p) \in \mathbb{R}^p$

à tout $j \in J$ est associé le I-descripteur : $(x_1^j, \dots, x_n^j) \in \mathbb{R}^n$.

Remarques

i) Si l'on remplace J par un autre ensemble J', on modifie la description des objets.

ii) Deux objets différents peuvent avoir la même description.

3.1. *b*) Si l'on note $\{\mathbf{b}_j \ / \ j \in]p]\}$ la base canonique de $\mathbb{R}^p$ et $\{\mathbf{a}^i \ / \ i \in]n]\}$ celle de $\mathbb{R}^n$:

$$\mathbf{x}_i = \sum_{j=1}^{p} x_i^j \mathbf{b}_j \qquad \mathbf{x}^j = \sum_{i=1}^{n} x_i^j \mathbf{a}^i.$$

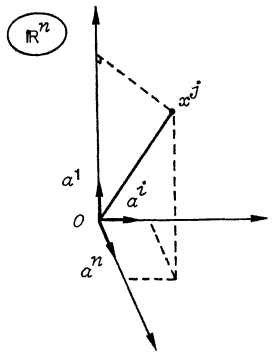

Figure 7.1

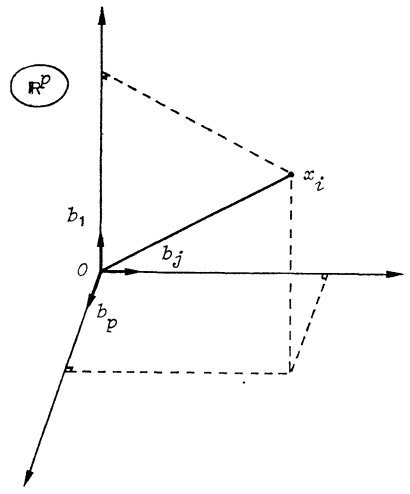

Figure 7.2

Figure 7. *Représentations du tableau X dans* $\mathbb{R}^n$ *et* $\mathbb{R}^p$

Enfin à chacune des descriptions d'objets, on affecte une masse m_i. On prendra $\sum_{i=1}^{n} m_i = 1$ et classiquement on suppose $m_i = \frac{1}{n}$.

(Cependant les résultats établis sont plus généraux : ils sont valables dès que $\sum_{i=1}^{n} m_i = 1$. C'est pourquoi la notation m_i est conservée.)

A chacun des éléments de l'ensemble I (resp. J) est associé un vecteur de $\mathbb{R}^p$ (resp. $\mathbb{R}^n$). A l'ensemble I est associé le nuage $\mathscr{N} = ((\mathbf{x}_i, m_i))_{i \in I}$ des descriptions pondérées.

Afin de répondre à certaines questions soulevées dans l'introduction sur la proximité des objets entre eux, des I-descripteurs entre eux, introduisons sur $\mathbb{R}^p$ et $\mathbb{R}^n$ des distances euclidiennes.

• Sur $\mathbb{R}^p$ définissons le produit *scalaire usuel* ; notons respectivement $(\mathbf{x}|\mathbf{x}')$ et $d(\mathbf{x},\mathbf{x}')$ le produit scalaire et la distance associée entre deux éléments x et x′ de $\mathbb{R}^p$:

$$d^2(\mathbf{x},\mathbf{x}') = \sum_{j=1}^{p} (x_j - x_j')^2 \qquad \text{avec} \qquad \mathbf{x} = \sum_{j=1}^{p} x_j \mathbf{b}_j \qquad \text{et} \qquad \mathbf{x}' = \sum_{j=1}^{p} x_j \mathbf{b}.$$

•• Sur $\mathbb{R}^n$ définissons le produit scalaire suivant :

$$(\mathbf{z}|\mathbf{z}')_c = \sum_{i=1}^{n} m_i z_i z_i' \qquad \text{avec} \qquad \mathbf{z} = \sum_{i=1}^{n} z_i a^i \text{ et } \mathbf{z}' = \sum_{i=1}^{n} z_i' \mathbf{a}^i.$$

Notons D la distance associée :

$$D^2(\mathbf{z},\mathbf{z}') = \sum_{i=1}^{n} m_i (z_i - z_i')^2.$$

Notons m_z et m_z les moyennes de z et z′ :

$$m_z = \sum_{i=1}^{n} m_i z_i \text{ et } m_{z'} = \sum_{i=1}^{n} m_i z_i' ;$$

$\tilde{\mathbf{z}} = \sum_{i=1}^{n} (z_i - m_z)\,\mathbf{a}^i$ et $\tilde{\mathbf{z}}' = \sum_{i=1}^{n} (z_i' - m_{z'})\,a^i$ les vecteurs centrés associés ;

alors : $(\tilde{\mathbf{z}}|\tilde{\mathbf{z}}')_c = \sum_{i=1}^{n} m_i (z_i - m_z)(z_i' - m_{z'})$

$= \mathit{Cov}\ (\mathbf{z},\mathbf{z}').$

De même : $(\tilde{\mathbf{z}}|\tilde{\mathbf{z}})_c = \mathit{Var}\ \mathbf{z}.$

La covariance de deux vecteurs z et z′ de $\mathbb{R}^n$ est égale au produit scalaire des vecteurs centrés $\tilde{\mathbf{z}}$ et $\tilde{\mathbf{z}}'$ associés.

3.2. *Résolution du problème*

3.2. *a) Meilleur ajustement par un sous-espace V de dimension r :* recherche des *r* droites vectorielles de l'approche du § 2.4.

On a vu que la résolution de ce problème nécessite l'étude des valeurs et vecteurs propres de la forme quadratique de dispersion :

$$S(\mathbf{u}) = \sum_{i=1}^{n} m_i (\mathbf{u}|\mathbf{x}_i - \mathbf{M})^2.$$

Le produit scalaire dans $\mathbb{R}^p$ ayant été précisé en (2.1.b).

La matrice Ψ est ici I_p, on peut déterminer :

3.2. *b) La matrice symétrique associée à la forme bilinéaire de dispersion s.*

Quelle est sa matrice relativement à la base canonique de $\mathbb{R}^p$? Le terme général de la matrice Σ associée à s est :

$$\forall\ i,\ j \in]p] \ :\ s(\mathbf{b}_j,\mathbf{b}_k) \ =\ \sum_{i=1}^{n} m_i(\mathbf{b}_j|\mathbf{x}_i - \mathbf{M})(\mathbf{b}_k|\mathbf{x}_i - \mathbf{M}),$$

$$\text{or} : \forall\ j \in]p],\ \forall\, i \in]n] \ :\ (\mathbf{b}_j|\mathbf{x}_i) \ =\ \mathbf{x}_i^j$$

$$: (\mathbf{b}_j|\mathbf{M}) \ =\ m_{x^j}$$

$$\text{donc } s(\mathbf{b}_j,\mathbf{b}_k) \ =\ \sum_{i=1}^{n} m_i(x_i^j - m_{x^j})(x_i^k - \mathrm{m}_{x^k})$$

$$s(\mathbf{b}_j,\mathbf{b}_k) \ =\ \textit{Cov}\ (\mathbf{x}^j,\mathbf{x}^k)$$

Donc la matrice Σ associée à la forme bilinéaire *de dispersion est la matrice des covariances des I-descripteurs*

$$\Sigma \ =\ (\textit{Cov}\ (\mathbf{x}^k,\mathbf{x}^j))_{1 \leqslant j,k \leqslant p}$$

Conclusion : les r droites vectorielles cherchées sont engendrées par des vecteurs propres unitaires $u_1, u_2, ..., u_r$ associés aux r plus grandes valeurs propres $\lambda_1 \geqslant \lambda_2 \geqslant ... \geqslant \lambda_r\ (\geqslant 0)$ de la matrice des covariances.

3.3. *Définitions* (3.3.1)

• *Axes factoriels :* $[u_k]$ est appelé *$k^{ième}$ axe factoriel* du nuage $\mathscr{N} \subset \mathbb{R}^p$ ou *$k^{ième}$* direction (ou axe) principale de dispersion.

• *Composantes principales :* les coordonnées par rapport à u_k des x_i de la famille $(x_i)_{i \in I}$, constituent une suite de n nombres réels, notée $c^k = (c_i^k)_{i \in I}$. On l'appelle *$k^{ième}$ composante principale* de $\mathscr{N}$.

On remarquera que $c^k = \sum_{i=1}^{n} c_i^k \mathbf{a}^i \in \mathbb{R}^n$.

Remarque : Une composante principale peut être regardée comme un I-descripteur non observé, combinaison linéaire des I-descripteurs initiaux.

Les r premières composantes principales fournissent une nouvelle description des objets.

Proposition (3.3.1) : Deux composantes principales distinctes c^j et c^k du nuage sont *non-corrélées* (bien que non-orthogonales).

Proposition (3.3.2) : La variance de la $k^{\text{ième}}$ composante principale c^k est égale à la valeur propre λ_k associée à u_k de s.

Preuve

Comme pour (3.2.b) on démontre que :

$$\forall\ j,k \in]p]\ \ s(u_j,u_k) = \sum_{i=1}^{n} m_i(c_i^j - m_{c^j})(c_i^k - m_{c^k})$$
$$= Cov\ (c^j,c^k).$$

Or, par définition, $s(u_j,u_k) = \lambda_j\ \delta_{jk}$;

donc : $Cov\ (c^j,c^k) = \lambda_j\ \delta_{jk}$.

On a donc bien résolu le problème de la recherche des I-descripteurs (composantes principales) non corrélés de variance décroissante.

3.4. *Analyse en composantes principales de tableaux centrés*

a) Soit le nuage $\tilde{\mathscr{N}} = ((\tilde{x}_i,m_i))_{i\in I}$, avec $x_i = \tilde{x}_i - M$ et $M = \sum_{i=1}^{n} m_i x_i$, obtenu par centrage [1] des I-descripteurs.

Puisque $\tilde{x}_i$ est l'image de x_i par la translation de vecteur $-M$ et que la translation est une isométrie, on a :

- pour tous les couples $(x_i,x_{i'})$ de vecteurs de $\mathscr{N}$ et $(\tilde{x}_i,\tilde{x}_{i'})$ de vecteurs de $\tilde{\mathscr{N}}$:

$$d(x_i,x_{i'}) = d(\tilde{x}_i,\tilde{x}_{i'})\ ;$$

- les formes quadratiques de dispersion associées à $\mathscr{N}$ et à $\tilde{\mathscr{N}}$ sont les mêmes

$$(\text{car } Cov(x_i,x_{i'}) = Cov(\tilde{x}_i,\tilde{x}_{i'})).$$

En conséquence :

- les directions des axes principaux,
- la qualité d'ajustement par les r premiers axes sont les mêmes pour $\mathscr{N}$ et $\tilde{\mathscr{N}}$.

En notant $c^k \in \mathbb{R}^n$ et $\tilde{c}^k \in \mathbb{R}^n$ deux composantes principales associées au $k^{\text{ième}}$ axe principal des nuages $\mathscr{N}$ et $\tilde{\mathscr{N}}'$:

- $c^k = \Pr_{[\delta]^\perp} c^k$ avec $\delta = \sum_{i=1}^{n} a^i$;
- $Cov\ (\tilde{c}^k,\tilde{c}^l) = Cov\ (c^k,c^l) = \delta_{kl}\ \lambda_k$;
- si $k \neq l$, $\tilde{c}^k$ *et* $\tilde{c}^l$ *sont orthogonaux.*

---

1. Pour l'interprétation géométrique de l'opération centrage (comme pour celle de réduction qui s'introduit dans 3.5) se reporter à [1] ou à une note à paraître dans cette revue.

b) Enfin remarquons que l'on peut analyser le tableau obtenu en centrant les descriptions des objets. Cette analyse aboutit en général à des résultats différents de ceux de $\mathscr{N}$.

3.5. *Représentations graphiques associées au nuage centré*

Comme dans l'exemple de la partie 1 les résultats de l'analyse sont fournis sous forme de tableaux et de graphiques :

— tableau des données, des moyennes, des écarts-types des I-descripteurs ;

— matrice des covariaces ou de corrélation des I-descripteurs (voir plus loin, § 3.6) ;

— valeurs propres et qualités de l'ajustement ;

— vecteurs propres ;

— composantes principales, projections des descriptions sur chacun des axes factoriels ;

— coordonnées des projections des I-descripteurs sur des vecteurs unitaires homothétiques à chacune des composantes principales ;

— qualité de la représentation des descriptions par chacun des axes factoriels ;

— qualité de la représentation des I-descripteurs par chacune des composantes principales.

3.5.1. *Dans* $\mathbb{R}^p$

A l'aide de programmes d'ordinateurs il est possible de représenter la projection du nuage $\mathscr{N}$ sur un plan, par exemple celui engendré par les deux premiers axes factoriels $[u_1]$, $[u_2]$. On peut se reporter à l'exemple de la partie 1.

Il est alors possible d'interpréter les proximités lues sur ce graphique. Pour deux descriptions x_i et x_i, *bien représentées* [1] (voir 2.5) sur ce plan, la distance lue est voisine de la distance réelle :

des points bien représentés proches sur le graphique correspondent à des descriptions proches dans l'espace ;

à des points bien représentés éloignés sur les graphiques correspondent à des descriptions éloignées dans l'espace.

On a ainsi une possibilité de constituer des classes de *descriptions proches* et d'apprécier les distances entre classes. Par exemple la classe C_1 est constituée des descriptions des comtés suivants : {*Athens, Belmont, Monroe, Noble...*}

la classe C_2 des descriptions des comtés : {*Clinton, Fayette, Greene, Warren...*}

la classe C_3 des descriptions des comtés : {*Allen, Auglaze, Hancock, Putnam, Sanousky, Seneca, Shelby, ...*}.

On projette aussi sur le plan $[u_1, u_2]$ les *directions* des vecteurs de bases $\{b_j / j \in]p]\}$ associées aux descripteurs. On obtient ainsi une représentation simultanée. En tenant compte des vecteurs b_j *bien représentés* [1] on peut juger du rôle que jouent ces vecteurs dans les oppositions ou les associations des classes entre elles ou des vecteur entre eux.

---

1. Le niveau à choisir dépend du problème.

En utilisant les vecteurs de base correspondant au Foin, Soja, Maïs et Blé

$q_{[u_1,u_2]}$ (foin) = 0.37 $\qquad$ $q_{[u_1,u_2]}$ (maïs) = 0.41

$q_{[u_1,u_2]}$ (soja) = 0.28 $\qquad$ $q_{[u_1,u_2]}$ (blé) = 0.31

et les projections des x_i sur ces vecteurs (voir Fig. 2 et Fig. 8), on remarque que :

C_1 est constituée de descriptions de comtés dont le pourcentage de surface cultivée en foin est très grand.

C_2 est constituée de descriptions de comtés dont le pourcentage de surface cultivée est grand en maïs mais faible en foin.

C_3 est constituée de descriptions de comtés dont le pourcentage cultivé est assez faible en foin et moyen en soja et blé.

Pour des vecteurs de base mal représentés (par exemple les petites graines $q_{[u_1,u_2]} = 0.04$) de telles pratiques sont dangereuses. Donnons un exemple : *GALLIA* et *HAMILTON* dont les pourcentages de superficie en petites graines sont respectivement la plus faible (0.00 %) et la plus forte (0.83 %) des 88 comtés se projettent sur l'axe associé aux petites graines en des points voisins [1] (voir Fig. 2 et Fig. 9).

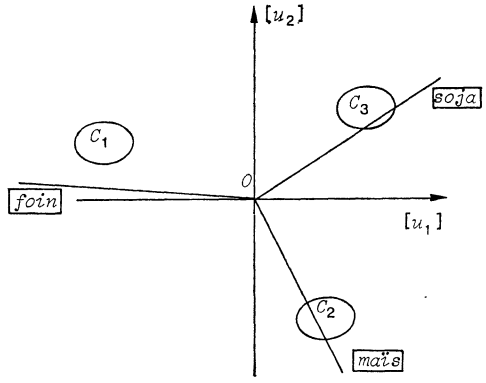

Figure 8. *Visualisation de classes (plan des deux premiers axes factoriels)*

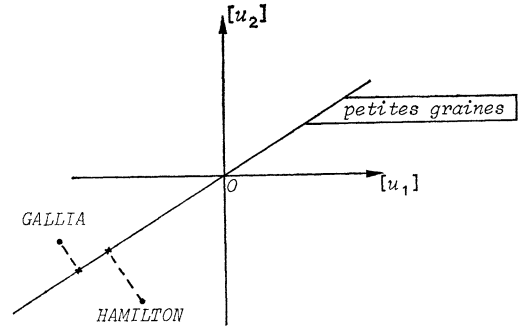

Figure 9. *Un danger de la visualisation*

3.5.2. *Dans* $\mathbb{R}^n$

Deux composantes principales, par exemple les deux premières c^1 et c^2 sont orthogonales et engendrent un plan vectoriel [2] sur lequel sont projetés les x^j qui sont les I-descripteurs.

On peut alors étudier et tenter d'interpréter, en tenant compte des qualités de représentation des x^j :

---

1. Le vecteur de base associé aux petites graines fait avec $[u_1, u_2]$ un angle dont le cosinus est voisin de 1.

2. Ce plan ne réalise pas le meilleur ajustement du nuage $((x^j, \frac{1}{p}))_{j \in J} \subset \mathbb{R}^n$.

a) les proximités entre les composantes principales et les I-descripteurs x^j (dans l'exemple le I-descripteur foin est proche de la 2e composante principale) ;

b) les proximités entre les I-descripteurs x^j.

Notons K la famille constituée des éléments de la famille $(x^j)_{j \in J}$ et ceux de la famille (c^1, c^2).

Puisque tous ces vecteurs sont centrés :

a) deux vecteurs $x \in K$ et $y \in K$ bien représentés et orthogonaux ont une covariance nulle et sont donc non-correlés. Voir par exemple sur le graphique 2 les projections des I-descripteurs avoine et soja ;

b) une sous-famille de I-descripteurs bien représentés, mutuellement proches constitue une classe de I-descripteurs. Si en plus les éléments de cette sous-famille sont fortement corrélés avec une composante principale, ils peuvent aider à l'interprétation de celle-ci.

3.6. *Analyse en composantes principales de tableaux centrés réduits*

Sur l'exemple géographique introductif on constate que les I-descripteurs peuvent avoir des variances très différentes : entre la variance du foin $(13.38)^2$ et les petites graines $(0.11)^2$ le rapport est supérieur à 15 600 (rapport des écarts-types 125). Cette différence influence profondément les résultats de l'analyse en C.P. :

L'analyse consistant à déterminer les directions sur lesquelles le nuage projeté est de dispersion maximum, la contribution des I-descripteurs petites graines et orge à la dispersion totale est très faible, puisque la dispersion du nuage est la somme des variances des I-descripteurs. Leur contribution à la détermination des premiers axes de dispersion sera négligeable.

Une pratique courante appelée *réduction* permet de rendre toutes égales à $\frac{1}{p}$ les contributions de chacun des I-descripteurs initiaux à la dispersion totale. A l'élément x^j (centré) de $\mathbb{R}^n$ on associe $y^j = x^j / \sqrt{Var\ x^j}$. Donc, quel que soit $j \in J$, le I-descripteur x^j est remplacé par le descripteur $y^j \in \mathbb{R}^n$.

En réduisant les vecteurs x^i on change le tableau X des données : on obtient un nouveau codage des mesures effectuées, donc d'autres descriptions des objets.

A l'objet $i \in I$ est associé le vecteur noté :

$$y_i = (x_i^1 / \sqrt{Var\ x^1}, \ldots, x_i^p / \sqrt{Var\ x^p}) \in \mathbb{R}^p.$$

Le nuage étudié est alors le nuage $\mathscr{N}' = \{(y_i, m_i) \ / \ i \in I\}$.

On est ramené à étudier le nuage $\mathscr{N}'$ défini ci-dessus la matrice des covariances associée à $\mathscr{N}'$ est la matrice des corrélations R associée à $\mathscr{N}$ donc les axes factoriels de $\mathscr{N}'$ sont les directions propres de R.

Représentations graphiques

Dans $\mathbb{R}^p$: Les règles de représentations et d'interprétations restent les mêmes que celles présentées en 3.5.1.

Dans $\mathbb{R}^n$: Tous les descripteurs ont même norme. Ils sont donc sur la sphère de centre 0 et de rayon 1.

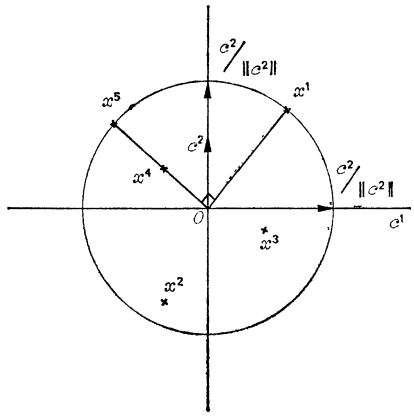

Figure 10. *Descripteurs centre-réduits : plan de deux premières composantes principales*

Sur le plan (C) des deux premières composantes principales, la sphère se projette sur la boule de centre 0 et de rayon 1. La coordonnée par rapport à $\frac{\mathbf{c}^1}{\|\mathbf{c}^1\|}$ (par exemple) d'un descripteur est égale au *coefficient de corrélation* entre $\mathbf{c}^1$ et ce descripteur.

Un point bien représenté par le plan (C) est sur le cercle ou au voisinage : $\mathbf{x}^1$, $\mathbf{x}^2$.

Un point près du centre est mal représenté : $\mathbf{x}^3$. Par ailleurs on *lit* sur la Fig. 10 l'égalité à zéro de la corrélation entre $\mathbf{x}^1$ et $\mathbf{x}^5$.

INDEX DES NOTATIONS

On notera :

— $\mathbb{N}$ l'ensemble des entiers naturels.

— $]r] = \{a \in \mathbb{N} \ / \ 1 \leqslant a \leqslant r\}$.

— $\mathbb{R}$ l'ensembe des nombres réels ;

— δ_{kl} (symbole de Kronecker) défini par $\delta_{kl} = \begin{cases} 0 \text{ si } k \neq l \\ 1 \text{ si } k = l. \end{cases}$

— p_V : projection orthogonale sur le sous-espace vectoriel V.

— $(x|y)$: produit scalaire des vecteurs x et y.

— $[u]$: droite vectorielle engendrée par le vecteur u.

— $[u_1, ..., u_r]$ sous-espace vectoriel engendré par les vecteurs $u_1, ..., u_r$.

— $V^{\perp}$ supplémentaire orthogonal de V.

— $|I|$ cardinal de l'ensemble I.

Conventions

— Deux espaces vectoriels $\mathbb{R}^p$ et $\mathbb{R}^n$ jouent un rôle important ; lorsqu'il est nécessaire d'indiquer les éléments de ces deux espaces on adopte la convention suivante :
indice haut pour $\mathbb{R}^n$ (ex. : $x^j \in \mathbb{R}^n$)
indice bas pour $\mathbb{R}^p$ (ex. : $y_i \in \mathbb{R}^p$).

— Deux ensembles finis I et J tels que $|I| = n$, $|J| = p$ jouent aussi un rôle important. On identifiera souvent I à $]n]$, J à $]p]$ et la notation suivante : $\sum_{i \in I}$ à $\sum_{i=1}^{n}$

— Les éléments de $\mathbb{R}^n$ ou $\mathbb{R}^p$ seront appelés indifféremment vecteurs ou points.

INDEX DES TERMES

ANNEXE I

Démonstration du théorème (2.4.1)

1. *Lemme préliminaire*

Soient V un sous-espace de dimension $r < n$ d'un espace euclidien E, [v] une droite de E qui n'est pas dans V, et $[v]^{+}$ l'orthogonal de v dans E ;

alors : $r - 1 \leqslant dim\ (V \cap [v]^{+}) \leqslant r. \qquad \leqslant r.$

En effet : $dim\ ([v]^{+} \cap V) = dim\ [v]^{+} + dim\ V - dim\ ([v]^{+} + V)$ (1).

a) Supposons $V \subset [v]^{+}$, alors $dim\ ([v]^{+}\ V) = n - 1$ et (1) entraîne : $dim\ ([v]^{+} \cap V) = r$;

b) Supposons $V \subset [v]^{+}$, alors $dim\ ([v]^{+} + V) = n$ et $dim\ ([v]^{+} \cap V) = r - 1$;

ce qui achève la démonstration.

2. *Démonstration du théorème*

Supposons que les vecteurs u_i sont unitaires.

Supposons que $\overset{r}{\underset{i=1}{\oplus}} [u_i] \not\subset V$ et montrons que dans ce cas V ne réalise par le meilleur ajustement, ce qui est contraire à la définition.

Soit $[u_l]$ la droite vectorielle telle que :

$$l = \underset{1 \leqslant i \leqslant r}{Min} \{i \ / \ [u_i] \not\subset V\},$$

et si $l > 1$ $\overset{l-1}{\underset{i=1}{\oplus}} [u_i] \subset V.$

Le lemme préliminaire entraîne : il existe un sous-espace vectoriel V_1 de V de dimension $r - 1$ tel que $V_1 \subset (V \cap [u_l]^{+})$ et si $l > 1$:

$$\overset{l-1}{\underset{i=1}{\oplus}} [u_i] \subset V_1.$$

Soit $\{v_1, v_2, ..., v_{r-1}\}$ une base orthonormale de V_1, qui contient $\{u_1, u_2, ..., u_{l-1}\}$ si $l > 1$.

Soit l'ensemble $\{v_1, v_2, ..., v_{r-1}, u_l\}$ et l'ensemble $\{v_1, v_2, ..., v_{r-1}, v\}$ dans lequel v complète l'ensemble $v_1, ..., v_{r-1}$ en une base orthonormée de V.

On a :

i) $v \neq u_l$ puisque $[u_l] \not\subset V$

ii) si $l > 1$, $\forall\ i \leqslant l - 1 : v \perp u_i$

iii) on appelle $W = [v_1, v_2, ..., v_{r-1}, u_l]$

$W \neq V$ d'après *i*) et d'autre part :

$$Disp\ \mathscr{M}_W = \sum_{i=1}^{r-1} Disp\ \mathscr{M}_{[v_i]} + Disp_{[u_l]} > Disp\ \mathscr{M}_V =$$

$$> \sum_{i=1}^{r-1} Disp\ \mathscr{M}_{[v_i]} + Disp\ \mathscr{M}_{[v]}$$

car $Disp\ \mathscr{M}_{[u_l]} > Disp\ \mathscr{M}_{[v]}$ d'après la définition, de u_l et d'après *ii*).

ANNEXE II

1. *Famille*

Définition (1.1) : Soient Λ et E deux ensembles. Une application de Λ dans E est souvent appelée une *famille d'éléments de E* ayant Λ comme ensemble d'*indices.*

On note : $l \mapsto x_l$ ($l \in \Lambda$, $x_l \in E$)

ou plus communément : $(x_l)_{l \in \Lambda}$

Remarque (1.1) : Il ne faut pas confondre la famille, c'est-à-dire une application, avec l'ensemble image de cette même application.

Exemple : $\Lambda =]4]$, $E = \mathbf{R}$.

$(x_i)_{i \in \Lambda}$ avec $x_1 = 3$, $x_2 = \frac{1}{3}$, $x_3 = -\sqrt{2}$, $x_4 = \frac{1}{3}$

$(y_i)_{i \in \Lambda}$ avec $y_1 = -\sqrt{2}$, $y_2 = \frac{1}{3}$, $y_3 = \frac{1}{3}$, $y_4 = 3$

ces deux familles sont différentes mais ont même ensemble image :

$$\{-\sqrt{2}, \frac{1}{3}, 3\}.$$

Remarque (1.2) : Si Λ est un ensemble fini, $(x_i)_{i \in \Lambda}$ est appelée *famille finie.*

Si $\Lambda = \mathbb{N}$ (resp. $\Lambda = P$, où P est un sous-ensemble fini de $\mathbb{N}$), $(x_i)_{i \in \Lambda}$ est appelée *suite* (resp. *suite finie*).

2. *Descripteur. Description :* Nous considérons deux ensembles finis I et J de cardinaux *n* et *p*.

Définition (2.1) : On appelle I-*descripteur* à valeurs dans $\mathbb{R}$ de l'ensemble I toute famille d'éléments de $\mathbb{R}$ ayant I pour ensemble d'indices.

Remarque (2.1) : L'ensemble I est équipotent à $]n]$, aussi très souvent nous identifierons un I-descripteur à une suite finie ayant $]n]$ pour ensemble d'indices.

Proposition (2.1) : L'ensemble $\mathcal{D}$ (I) des I-descripteurs est un sous-espace vectoriel pour les lois usuelles d'addition et de multiplication par un scalaire.

• Soit $\Delta = (\delta_j)_{j \in J}$, une famille finie de I-descripteurs.

Définition (2.2) : On appelle *description* de $i \in I$ associée à Δ (ou plus simplement description de $i \in I$) la famille $(\delta_j(i))_{j \in J}$.

Exemple : * I : ensemble des comtés de l'Ohio.

* I-descripteur : une suite de 88 nombres réels associée à une céréale.

* J : ensemble des noms des céréales.

* description : une suite de 7 nombres réels associée à chacun des comtés.

(*à suivre*)

Nous remercions Hélène Geroyannis et Josiane Leconte pour la réalisation matérielle du texte, le calibrage et l'exécution des graphiques.

BIBLIOGRAPHIE

[1] Anderson, R. C., *An introduction to multivariate analysis*, New York, J. Wiley, 1958.

[2] Barbut, M., *Mathématiques des Sciences humaines*, 2 t., Paris, Presses Universitaires de France, 1967.

[3] Benzécri, J.-P., *Leçons sur l'analyse statistique des données multidimensionnelles*, ISUP, Paris, 1969, Ronéo.

[4] Dempster, A. P., *Elements of continuous multivariate analysis*, New York, Addison Wesley, 1969.

[5] Deniau, C., Leroux, B., Oppenheim, G., *Deux méthodes d'analyses factorielles*, Actes du colloque « Analyse des données en Architecture et Urbanisme », Institut de l'Environnement, 1972.

[6] Gnanadesikan, R., Kettering, J. R., « Robust estimates, residuals, and outlier detection with multireponse data », *Biometrika*, mars 1972, vol. 28, n° 1, pp. 81-124.

[7] King, L., *Statistical analysis in geography*, New York, Prentice-Hall, 1969.

[8] Rao, C. R., « The use and interpretation of principal component analysis in applied research », *Sankhya*, A, 26, 1964, pp. 329-358.

[9] Weaver, J. C., « Crop combination regions in the Middle West », *Geographical review*, vol. 44, 1954, pp. 175-200.

[10] Weaver, J. C., « Changing patterns of cropland use in the Middle West », *Economic geography*, vol. 30, 1954, pp. 1-47.