REVUE DE STATISTIQUE APPLIQUÉE

JEAN-PIERRE NAKACHE

Influence du codage des données en analyse factorielle des correspondances étude d'un exemple pratique médical

Revue de statistique appliquée, tome 21, n° 2 (1973), p. 57-70

<http://www.numdam.org/item?id=RSA_1973__21_2_57_0>

INFLUENCE DU CODAGE DES DONNÉES EN ANALYSE FACTORIELLE DES CORRESPONDANCES ÉTUDE D'UN EXEMPLE PRATIQUE MÉDICAL

Jean-Pierre NAKACHE
Groupe de recherche U 88
C.H.U. Pitié Salpétrière (Service du Professeur Grémy)

I – INTRODUCTION

L'analyse factorielle des correspondances (Réf. 2, 2', 2'', 3, 4, 7, 9), dont le principe est exposé succinctement en annexe, est une méthode d'analyse descriptive multidimentionnelle qui s'applique rigoureusement à des tableaux de contingence à n lignes et p colonnes. Les lignes représentent en général les "individus" et les colonnes, les "caractères" ou paramètres mesurés sur chaque individu. Dans la case (i, j) d'un tableau de contingence se trouve le nombre *positif* d'occurences du couple (i, j). Dans un tel tableau, les lignes et les colonnes jouent des rôles symétriques. L'analyse factorielle des correspondances conduit à la *représentation simultanée* des points des deux ensembles – ensemble des individus, ensemble des caractères – dans un même espace de dimension restreinte, tout en respectant au mieux les "proximités" ou distances entre les points des deux ensembles. L'analyse du double nuage des points-individus et des points-caractères repéré dans les plans principaux significatifs engendrés par les axes d'élongation maximale du nuage, permet de définir des classes d'individus homogènes, si elles existent, de caractériser ces classes, de déterminer les liaisons entre caractères et de donner une interprétation des différents axes, si elle est jugée utile.

Les données médicales, et bien d'autres données, sont souvent de natures différentes : quantitatives, qualitatives à deux modalités, qualitatives à plus de deux modalités. Dans ce cas, les lignes et les colonnes du tableau de données à analyser ne jouent pas des rôles symétriques. La pondération par colonne a toujours un sens, mais, il n'est pas de même de la pondération par ligne. En effet la somme, pour un malade donné, de ses mesures sur les différents paramètres hétérogènes n'a pas beaucoup de signification, ce qui diminue l'intérêt de la distance du chi-2.

Malgré l'hétérogénéïté de telles données, on peut avoir recours à l'analyse factorielle des correspondances mais les résultats obtenus se révèlent en pratique assez pauvres. Les représentations graphiques issues d'une telle analyse semblent difficilement interprétables : l'information extraite est évidente ou sans grande importance, comme on le verra dans l'exemple pratique.

Il convient donc d'effectuer sur ce genre de données un codage binaire en vue d'utiliser l'analyse factorielle des correspondances dans les meilleures conditions.

Les tableaux de données, sont par ce codage, transformées en tableaux de données binaires (présence-absence) qui, eux, respectent les propriétés des tableaux de contingence dont ils sont des cas particuliers.

Dans le présent article, on utilisera des notions générales de la théorie de l'information pour mettre en évidence la perte d'information entraînée par un codage binaire de données hétérogènes. On indiquera ensuite certains critères pratiques permettant d'optimiser, dans le cas général, le choix du codage en vue de l'utilisation de l'analyse factorielle des correspondances ; on étudiera enfin le codage défini dans l'exemple pratique et son influence sur les résultats obtenus.

2 – CODAGE DES DONNEES

La décomposition de paramètres en caractères booléens peut s'accompagner d'une perte d'information. Le codage binaire des données est d'autant plus satisfaisant que cette perte d'information est minimum. Pour évaluer cette perte d'information on peut utiliser la notion d'information de SHANNON (Réf. 11, 5, 12).

Eléments de la théorie de l'information

Ces éléments sont en partie extraits de la Thèse de M. DELABRE (Réf. 5).

On considère un système E pouvant se trouver dans l'un des n états suivants : $E_1, E_2, \ldots, E_n$. On suppose que le système E ne peut pas se trouver simultanément dans plusieurs états, c'est-à-dire que les états sont des évènements incompatibles.

Soit alors P_i la probabilité pour que E se trouve dans l'état E_i.

$$P_i = \text{Prob}\{E = E_i\}$$

$$0 \leqslant P_i \leqslant 1 \qquad \text{et} \qquad \sum_{i=1}^{n} P_i = 1$$

2.1 – Incertitude

G. SHANNON (Réf. 11, 12) définit l'incertitude sur l'état du système E, *avant* toute observation, par la quantité

$$H(E) = - \sum_{i=1}^{n} P_i \log_2 P_i$$

H(E) est aussi appelé *entropie d'information de E*

2.2 – Information moyenne

L'information moyenne I(E) apportée par l'observation d'un système E est, par définition, égale à l'incertitude sur l'état du système avant toute observation.

$$I(E) = H(E)$$

2.3 – Perte d'information entraînée par la décomposition d'un paramètre en caractère booléns (Réf. 5).

Décomposition d'un paramètre quantitatif en variable qualitative à k modalités

soit $\{E_1, E_2, \ldots, E_n\}$ l'ensemble des n valeurs prises par un paramètre quantitatif E et C une partition de E en k classes

$$C_r = \{E_i / b_{r-1} \leqslant E_i < b_r\} \qquad r = 1, 2, \ldots, k$$

b_{r-1} et b_r sont les bornes de la classe C_r

C représente une variable qualitative à k modalités.

Soit, d'autre part, $f_i (i = 1, 2, \ldots, n)$ la fréquence d'occurence de la valeur E_i de E et f'_r $(r = 1, 2, \ldots, k)$ la fréquence d'occurence de la classe C_r de C.

On suppose la taille de l'échantillon à étudier suffisamment grande pour que l'on puisse appliquer la loi des grands nombres.

L'information de E est :

$$I(E) = \sum_{i=1}^{n} f_i \log_2 \frac{1}{f_i}$$

L'information transmise par le codage de E par C est :

$$I_C = \sum_{j \in I_r} f'_r \log_2 \frac{1}{f'_r}$$

ou

$$f'_r = \sum_{j \in I_r} f'_j$$

et I_r est l'ensemble des indices des valeurs E_i contenues dans la classe C_r.

La perte d'information entraînée par le passage de E à C est :

$$PI = I(E) - I(C) = H(E) - H(C)$$

soit :

$$PI = \sum_{i=1}^{n} f_i \log_2 \frac{1}{f_i} - \sum_{r=1}^{k} f'_r \log_2 \frac{1}{f'_r}$$

$$PI = \sum_{r=1}^{k} \left(\sum_{j \in I_r} f_j \log_2 \frac{1}{f_j} \right) - \sum_{r=1}^{k} \left(\sum_{j \in I_r} f_j \right) \log_2 \frac{1}{\sum_{j \in I_r} f_j}$$

$$PI = \sum_{r=1}^{k} \sum_{j \in I_r} f_j \left(\log_2 \frac{1}{f_j} - \log_2 \frac{1}{\sum_{j \in I_r} f_j} \right)$$

Comme la fonction $f(x) = \log_2 x$ est monotone croissante et que $f_j < \sum_{j \in I_r} f_j$, la perte d'information PI dans la transmission $E \rightarrow C$ est positive. *Il y a bien une perte d'information au cours du codage d'un paramètre quantitatif en un paramètre qualitatif à k modalités.*

Si $\mathcal{P}$ est l'ensemble des partitions possibles de E, on choisira la partition C_o telle que

$$I_{C_o} = \sup_{c \in \mathcal{P}} I_C$$

Autrement dit, on choisira C_o telle que la perte d'information entraînée par le codage de E en C_o soit minimale.

Proposition I

La perte d'information résultant de la transformation d'un paramètre quantitatif E en un caractère qualitatif C à k modalités est minimale lorsque *les différentes modalités sont équiprobables et le nombre de modalités maximal.*

On a remarqué plus haut que $P(I) = H(E) - H(C)$. Ainsi, rendre minimum la quantité P(I) revient à rendre maximum la quatité H(C)

$$H(C) = - \sum_{r=1}^{k} f'_r \log_2 f'_r$$

En utilisant les propriétés de la fonction connexe $f'_r \log_2 f'_r$ $(0 \leqslant f'_r \leqslant 1)$ et le fait que $\sum_{r=1}^{k} f'_r = 1$, on démontre (Réf. 5) que

$$H(C) \leqslant \log_2 k$$

Si, d'autre part, dans l'expression de H(C) on pose $f'_r = \frac{1}{k}$, on obtient $H(C) = \log_2 k$

Par conséquent H(C) atteint son maximum $\log_2 k$ pour

$$f'_r = \frac{1}{k} \text{ (pour } r = 1, 2, \ldots, k)$$

c'est-à-dire si les fréquences d'occurences des différentes classes sont égales. De plus, H(C) est d'autant plus grande que k est grand (c.q.f.d.).

2.5 – Décomposition d'un paramètre qualitatif à k modalités en k caractères booléens

Proposition 2

Le passage d'un paramètre qualitatif E à k modalités à k variables booléenes $B_1, B_2, \ldots, B_k$ *n'entraîne aucune perte d'information.*

Le passage de E à $(B_1, B_2 \ldots, B_k)$ peut être considéré comme la transmission de l'information d'un système E (à k états E_r de probabilités P_r) à un système $B_1 \times B_2 \times \ldots \times B_k$ produit de k systèmes à 2 états (+) et (−) tels que :

$$\text{Prob}\{B_r^{(+)}\} = P_s \quad \text{si} \quad r = s$$

$$\text{Prob}\{B_r^{(+)} \cap B_s^{(+)}\} = 0 \text{ si } r \neq s$$

Il en résulte que Prob $(B_1, B_2, \ldots, B_k)$ est nul s'il existe plus d'un système B_r dans l'état (+) ou si tous les systèmes sont dans l'état (−) et vaut P_r si B_r est dans l'état (+)

Par conséquent :

$$H(B_1, B_2, \ldots, B_k) = -\sum_{r=1}^{k} P_r \log_2 P_r = H(E)$$

L'information transmise par ce codage n'est donc pas perturbée (c.q.f.d.).

2.6 – Décomposition d'un paramètre qualitatif en deux variables booléenes

Proposition 3

Le passage d'un paramètre qualitatif E à 2 variables booléenes $E^{(+)}$ et $E^{(-)}$ n'entraîne aucune perte d'information.

Dans ce cas, le codage binaire $(E^{(+)}, E^{(-)})$ de E correspond à la dualité présence - absence et l'information transmise par le codage est

$$I = \text{Prob}(E^{(+)} \log_2 \frac{1}{\text{Prob}(E^{(+)})} + \text{Prob}(E^{(-)} \log_2 \frac{1}{\text{Prob}(E^{(-)})}$$

c'est-à-dire l'information sur le paramètre E lui-même.

En conclusion

Le codage binaire de paramètres qualitatifs par l'introduction de deux variables booléenes (présence/absence) ou de paramètres qualitatifs à k modalités par l'introduction de k variables booléens ne produit pas de perte d'information.

Le codage d'un paramètre quantitatif par k variables booléenes introduit une perte d'information que l'on minimise en choisissant la partition de l'ensemble des valeurs de ce paramètre en un nombre maximum de classes, chacune ayant la même fréquence d'occurence.

2.7 – Critères pratiques d'optimisation du choix d'un codage binaire de données.

Cas d'un paramètre quantitatif

On définira, après examen de l'histogramme du paramètre et en accord avec le spécialiste de l'étude qui connait bien les paramètres de son étude, des classes représentatives ou significatives. Ces classes devront contenir autant que possible le même effectif ; elles constitueront de nouvelles variables : variables boolénnes.

Cas d'un paramètre qualitatif à k modalités

Si les différentes modalités du paramètre *ne sont pas ordonnées* on remplacera tout simplement chacune des k modalités par une variable binaire (en présence/absence).

Si les modalités du paramètre sont *ordonnées* on pourra aussi remplacer chacune des modalités par une variable binaire ou, si c'est possible, grouper ces différentes modalités, de façon convenable, en classes et remplacer chaque classe par une variable booléenne. On pourra avoir recours à cette solution si le paramètre peut s'y prêter et surtout dans le cas où on veut limiter la multiplication des variables obtenues après codage.

Cas d'un paramètre qualitatif

Un paramètre qualitatif (présence/absence) sera dédoublé si on veut donner dans l'analyse autant de poids à l'absence du paramètre qu'à sa présence. Dans ce cas, on le remplacera par deux variables booléenes : absence (0,1) et presence (0,1). Si, dans l'analyse à effectuer, la présence d'un paramètre qualitatif à deux modalités est plus importante, *ou plus informative,* que son absence, on ne transformera pas ce paramètre.

3 – EXEMPLE MEDICAL

Les données de cet exemple proviennent du service d'Allergologie du Professeur HALPERN (Hôpital Broussais). Elles concernent un groupe de 67 asthmatiques caractérisés par la positivité du test au pollen des graminés. Il s'agit, dans cette étude, d'étudier l'influence de l'allergie pollinique et de l'infection sur le développement de la maladie asthmatique en vue de confirmer ou d'infirmer l'impression clinique et le choix thérapeutique.

Les paramètres mesurés sur chacun des individus sont de nature différente : qualitatives à deux modalités comme le sexe (masculin, féminin), qualitatives à plusieurs modalités *ordonnées* comme l'intensité de l'asthme (léger, moyen, fort ou invalidant), ou *non ordonnées* comme les prélèvements bactériologiques (pharyngé, expectoration, naso-sinusien) et quantitatives comme l'âge (en années).

A partir de ces données qui constituent le codage initial, on a essayé de définir un codage binaire en accord avec le médecin, en tenant compte des caractéristiques de chacun des paramètres et en évitant d'introduire une part d'arbitraire ou des hypothèses supplémentaires.

On a procédé de la manière suivante :

Pour les paramètres quantitatifs

Si on introduit dans une A. F. C. des paramètres quantitatifs à côté de variables binaires, ces dernières interviennent très peu car leur poids, dans l'ana-

lyse, est nul ou presque. Pour éviter d'introduire des déséquilibres injustifiés entre les poids des différents individus, on a tout intérêt à transformer les variables quantitatives. Dans l'exemple médical, la seule variable quantitative, âge au début du traitement, exprimée en années a été convenablement décomposée en 4 classes correspondant à 4 variables binaires : "AG1" (de 1 à 2 ans), "AG2" (de 3 à 8 ans), "AG3" (de 9 à 12 ans) et "AG4" (de plus de 12 ans). Ainsi un malade âgé de 10 ans présente les valeurs 0 pour "AG1", 0 pour "AG2", 1 pour "AG3" et 0 pour ',AG4".

Pour les paramètres qualitatifs à deux modalités

Si, dans l'A.F.C, une variable qualitative prend la valeur 1 (ou 0) pour la première modalité (absence) et la valeur 2 (ou 1) pour la deuxième modalité (présence) on donne plus d'importance à la deuxième modalité qu'à la première ce qui peut entraîner une disymétrie dans le tableau des données. Cet effet se produit si l'absence du paramètre est aussi informative que la présence. Par exemple, en ce qui concerne la variable qualitative "dermite atopique", il est aussi important de savoir si un malade présente une dermite atopique ou s'il ne la présente pas, quand on étudie la maladie asthmatique. Ce genre de variable qualitative a été systématiquement dédoublé en une variable présence (1-oui, 0-non) et une variable absence (1-oui, 0-non). On met ainsi clairement en évidence, dans l'analyse, l'influence respective de chacune des deux modalités.

Si, par contre, la présence d'une variable (présence, absence) est plus importante que son absence, le dédoublement de cette variable est alors superflu. C'est le cas, par exemple dans l'étude de l'asthme de la variable "hypotrophie" qui prend la valeur 1 si le malade présente une hypotrophie et 0 sinon.

Pour les paramètres qualitatifs à plus de 2 modalités

On a considéré dans l'exemple médical, une variable binaire (présence/absence) pour chaque modalité de ces paramètres. Par exemple le paramètre"intensité de l'asthme" a été décomposé en 3 variables binaires :

asthme léger = "ASL" = 1 si le malade présente un asthme léger, 0 sinon
asthme moyen = "ASM" = 1 si le malade présente un asthme moyen, 0 sinon
asthme fort = "ASF" = 1 si le malade présente un asthme moyen, 0 sinon

3. 1 – Codage des données

Pour étudier l'influence du codage en analyse factorielle des correspondances, on a utilisé, dans une première A.F.C., les paramètres de l'étude codées initialement, et, dans une deuxième A.F.C., les nouvelles variables résultant du codage binaire.

La liste complète des paramètres des deux codages figure dans le tableau des données codées. Les sigles des paramètres des deux codages permettent d'identifier ces paramètres dans les représentations graphiques des deux analyses factorielles des correspondances.

3.2 – Résultats des deux A.F.C. et interprétation

Les deux graphiques I et II représentent la projection du double nuage (points-malades, points-paramètres) sur le meilleur plan principal engendré par les deux premiers axes factoriels obtenus en effectuant une analyse factorielle des correspondances, d'une part, sur les 17 paramètres issus du codage initial, et, d'autre part, sur les 39 paramètres résultant du codage binaire.

L'examen du graphique I, résultant d'une A.F.C. effectuée à partir des données initiales, montre que la description des données est assez pauvre. Dans le but de tirer une information concernant la gravité de l'asthme, on a représenté, sur ce graphique, les malades différemment suivant l'intensité de leur asthme mais aucune différenciation entre les trois types d'asthme n'a pu être clairement mise en évidence. Il en a été de même pour les autres paramètres intéressants de l'étude à savoir l'évolution de la maladie et le facteur saisonnier. Néanmoins la proximité entre ces trois paramètres IAS, DEV et FSA semble indiquer leur liaison ou corrélation qui reste à confirmer par une autre analyse étant donné leur position particulière dans la zone proche du centre de gravité.

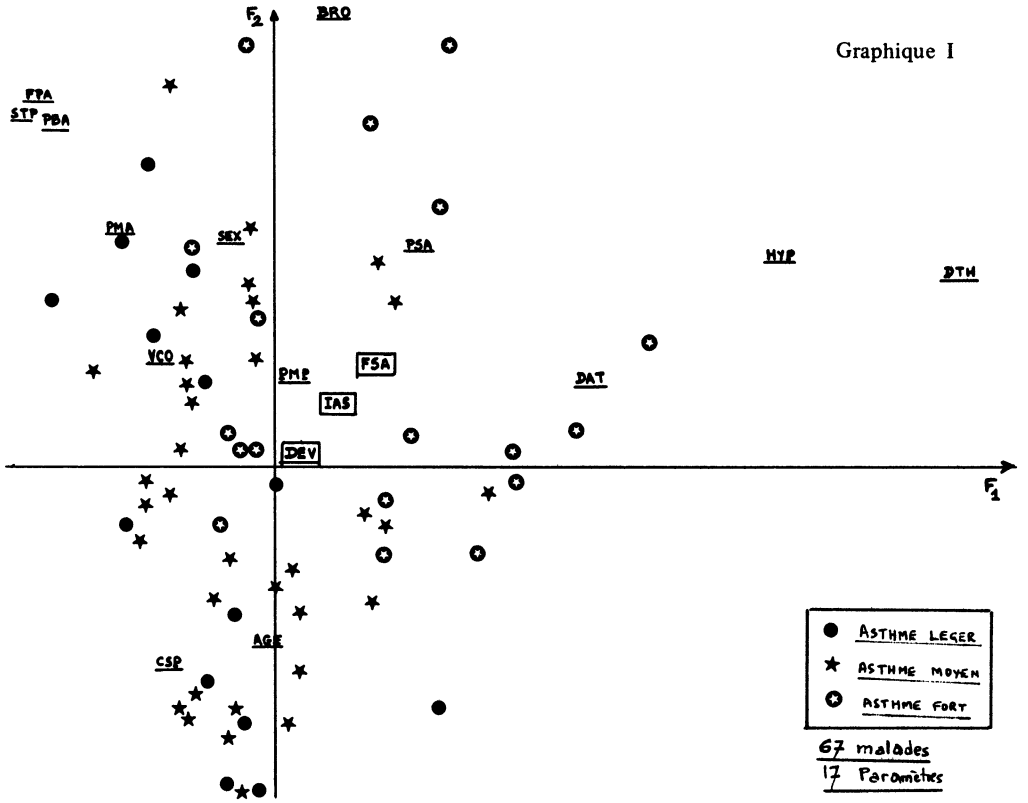
Ce sont les résultats plutôt inintéressants de cette première analyse qui nous ont amené à envisager la nouvelle codification (codage binaire) des données de l'étude pour essayer d'obtenir une description plus intéressante des données.

Le graphique II, résultant d'une A.F.C. effectuée à partir des nouvelles variables binaires, semble manifestement plus riche en information. A l'inverse du graphique I, il permet de mettre en évidence trois groupes caractéristiques de malades correspondant à trois formes cliniques d'asthme. Le premier groupe, situé au bas du graphique II, rassemble les malades qui présentent un asthme léger (ASL). Cette forme d'asthme apparaît au printemps ou en été (FSE) et évolue depuis 5 ans environ (E-1, E 15). Elle est accompagnée d'un coryza spasmodique (CS2). Les malades de ce groupe ne présentent pas de dermite atopique (DAA) et leurs tests allergologiques sont plutôt négatifs (PUA, POA). Enfin ce type d'asthme est davantage celui de la fille (SFE). Le deuxième groupe, situé en haut et à gauche du graphique II contient les malades ayant un asthme fort (ASF) durant toute l'année (FSP) et évoluant depuis la prime enfance (EPE). Cette forme d'asthme est accompagnée d'une hypotrophie (HYP), d'une déformation thoracique (DTH), d'une dermite atropique (DA1, DA2) et une absence de coryza spasmodique (CSA).

Le troisième groupe, situé au milieu et à droite du graphique II, est un peu moins net que les précédents. Il rassemble les malades ayant un asthme d'intensité moyenne (ASM), forme d'asthme à mi-chemin entre l'asthme fort et l'asthme léger. Cette forme d'asthme est à prédominance automnohivernale (FSA) évoluant depuis plus de 5 ans (E + 5). Elle est accompagnée de rhinopharyngites (RHI) avec une flore pharyngée pathogène (PHA). Les malades de ce groupe ont été atteints d'un coryza spasmodique (CSI).

Les bronchites (BRO), la positivité des tests allergologiques semblent être le fait de ce type d'asthme et de l'asthme fort.

Graphique I



Graphique II

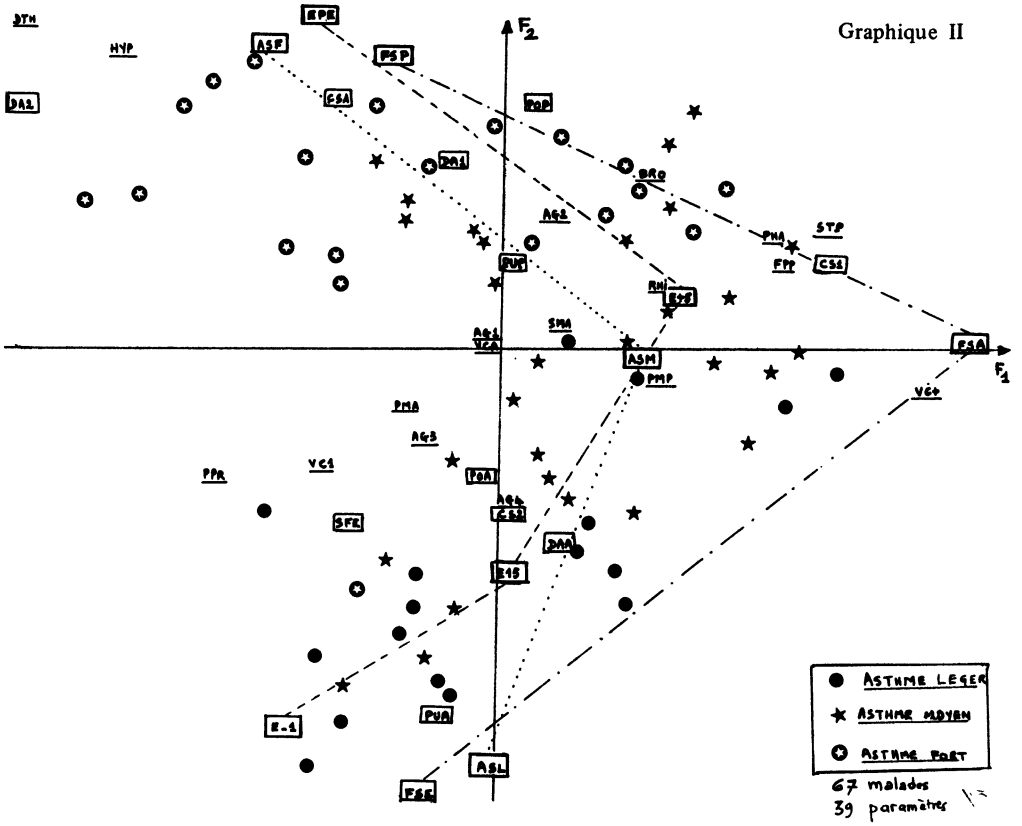


TABLEAU DES DONNEES CODEES

Paramètres de l'étude	Sigle	Codage initial	Sigle	Codage binaire
Age au début dutraitement en années)	AGE	(en classe d'âge 1 à 2 ans	AG1	1 - oui 0 - non
		3 à 8 ans	AG2	1 - 0 -
		9 à 12 ans	AG3	1 - 0 -
		plus de 12 ans	AG2	2 - 0 -
Sexe	SEX	1 - Masculin	SMA	1 - 0 -
		2 - Féminin	SFE	1 - 0 -
Intensité de l'asthme	IAS	1 - Asthme léger	ASL	1 - 0 -
		2 - Asthme moyen	ASM	1 - 0 -
		3 - Asthme fort	ASF	1 - 0 -
Facteur saisonnier	FSA	1 - Printemps-été	FSE	1 - 0 -
		2 - Automne-hiver	FSA	1 - 0 -
		3 - Perannuel	FSP	1 - 0 -
Durée d'évolution de la maladie	DEV	1 - moins d'un an	E - 1	1 - 0 -
		2 de 1 à 5 ans	E15	1 - 0 -
		3 - plus de 5 ans	E+5	1 - 0 -
		4 - depuis la prime enfance	EPE	1 - 0 -
Infection rhino-pharyngée	RHI	0 Absence 1 - Présence	RHI	1 - 0 -
Infection bronchique	BRO	0 - Absence 1 - Présence	BRO	1 - 0 -
Déformation thoracique	DTH	0 - Absence 1 - Présence	DTH	1 - 0 -
Hypotrophie	HYP	0 - Absence 1 - Présence	HYP	1 - 0 -
Dermite atopique	DAT	0 - Absence	DAA	1 - 0 -
		1 - ayant existé	DA1	1 - 0 -
		2 - Existant actuellement	DA2	2 - 0 -
Coryza spasmodique	CSP	0 - absence	CSA	1 - 0 -
		1 - ayant existé	CS1	1 - 0 -
		2 - existant actuellement	CS2	1 - 0 -
Antécédents all. familiaux				
a) verticaux, collatéraux	VCO	0 - absence	VCA	1 - 0 -
		1 - un sujet	VC1	1 - 0 -
		2 - plusieurs sujets	VC+	1 - 0 -
b) paternels, maternels	PMA	0 - absence	PMA	1 - 0 -
		1 - présence	PMP	1 - 0 -
Prélèvements bactériologiques	PBA	0 - aucun	PPR	1 - 0 -
		1 - pharyngé	PHA	1 - 0 -
Flore pathogène	FPA	0 absence	FPA	1 - 0
		1 - présence	FPP	1 - 0 -
Staphylocoques	STP	0 - absence	STA	1 - 0 -
		1 - présence	STP	1 - 0
Allergie aux				
a) poussières, plumes	PMP	0 - absence	PUA	1 - 0 -
		1 - présence	PUP	1 - 0 -
b) poils & squames d'animaux	PSA	0 - absence	POA	1 - 0 -
		1 - présence	POP	1 - 0 -

Les trois courbes tracées sur le graphique II ont, dans l'ensemble, la même direction ce qui confirme la corrélation des trois paramètres : intensité de l'asthme (IAS), durée d'évolution de la (DEV) maladie et facteur saisonnier (FSA).

3.3 – Remarque sur la significativité du graphique II

Les pourcentages d'inertie totale expliquée par les cinq premiers axes factoriels obtenus en effectuant l'A.F.C. sur les données résultant du codage binaire sont : 14 % ; 13,20 % ; 9 % ; 7,4 % et 6,5 %.

On peut constater une différence assez sensible entre les pourcentages d'inertie expliquée par les 2^e et 3_e axes factoriels et qu'ensuite la décroissance du pourcentage est plus lente. On s'est contenté, pour cette raison, de n'interpréter que le meilleur plan factoriel (F_1 , F_2).

Les proximités des points du double nuage ont été certainement déformées par projection sur le plan (F_1 , F_2) puisque l'inertie expliquée par ce plan ne représente que 27,20 % de l'inertie totale. La représentation graphique des points sur les axes factoriels F_3, F_4, F_5 aurait été souhaitable mais on a préféré s'en tenir aux renseignements donnés par le plan (F_1 , F_2) et essayé plutot d'utiliser d'autres analyses multidimensionnelles (Réf. 1, 4, 8) pour les confirmer. On a, dans ce but, effectué une classification arborescente sur les malades de l'étude en utilisant la distance du chi-2 pour définir les proximités entre points-malades de l'étude. On a ainsi retrouvé, à peu de choses près, la typologie mise en évidence dans l'A.F.C. On a aussi utilisé l'analyse discriminante pour déterminer parmi les paramètres de l'étude (définis dans le codage initial), ceux qui différencient au mieux les deux groupes bien distincts : asthme léger et asthme fort. On a eu la satisfaction de retrouver les caractères discriminants mis en évidence dans l'A.F.C. à savoir le facteur saisonnier, la dermite atopique, le coryza spasmodique, la durée d'évolution de la maladie. Enfin les résultats d'une régression multiple ont montré que les paramètres expliquant au mieux la variable "intensité de l'asthme" sont par ordre d'importance : caractère saisonnier, durée d'évolution de la maladie, dermite atopique... L'interprétation de la représentation graphique de l'analyse factorielle des correspondances effectuée à partir des variables résultant du codage binaire s'est ainsi trouvée confirmée par ces analyses plus fines.

4 – ANNEXE

Résumé de la méthologie de l'analyse factorielle des correspondances.

L' analyse factorielle des correspondances mise au point par J.P. BENZECRI (Réf. 2, 2', 2'', 3) est une analyse en composantes principales effectuée sur un tableau de contingence (n , p) appelé aussi tableau de correspondance. Elle ne diffère de cette dernière que par le choix de la distance utilisée. L'analyse factorielle des correspondances est basée sur la distance du chi-2 (Réf. 3, 9).

De tels tableaux d'observations peuvent tout aussi bien être traités par l'analyse en composantes principales classique mais il est plus naturel de définir une méthode qui tienne compte du caractère probabiliste de ce type de données et de la symétrie des rôles des lignes et des colonnes mis en correspondance.

Dans l'analyse factorielle des correspondances, les données initiales n_{ij} du tableau sont transformées en fréquences conditionnelles

$$y_{ij} = \frac{P_{ij}}{P_{i.}} \quad \text{où} \quad P_{ij} = \frac{n_{ij}}{\sum_i \sum_j n_{ij}}$$

est la fréquence relative du couple (i , j) et $P_{i.}$ est la fréquence marginale de la ième ligne du tableau $\left(P_{i.} = \sum_j P_{ij}\right)$, et ceci, dans le but de rendre comparables des différents profils des individus ou lignes du tableau. D'autre part, pour éviter de faire jouer à l'une des colonnes ou caractères un rôle excessif, on pondère les fréquences conditionnelles y_{ij} par la quantité $\sqrt{P_{.j}}$ ou $P_{.j}$ est la fréquence marginale de la j.ième colonne.

Si on s'intéresse à l'analyse des n points-individus dans l'espace R^P, on effectue alors une analyse en compostantes principales du nuage N_I des points x_i repérés dans R^P muni de la métrique unité ($M = I_p$), de coordonnées

$$x_{ij} = \frac{y_{ij}}{\sqrt{P_{.j}}}$$

(j = 1, 2, . . . , P) et affectés des poids respectifs P_i.

Cette analyse en composantes principales classique effectuée à partir des x_i est équivalente à l'analyse en composantes principales du nuage des points y_i repérés dans l'espace R^P muni de la métrique dite du chi.2, de coordonnées $y_{ij} = \frac{P_{ij}}{P_{i.}}$

Dans R^P muni de la métrique du chi-2, la distance entre deux individus "i" et "k" du nuage est '

$$d^2(i, k) = \sum_{j=1}^{P} \frac{1}{P_{.j}} \left(\frac{P_{ij}}{P_{i.}} - \frac{P_{Kj}}{P_{k.}}\right)^2$$

C'est la *distance du chi-2 (Réf. 3, 9) associée à* P_j . généralisation de la distance du chi.2 classique entre une distribution théorique et une distribution empirique. Les colonnes et lignes d'un tableau de contingence étant symétriques, on peut aussi effectuer une analyse en composantes principales du nuage N_J des p points-caractères dans l'espace R^n engendré par les vecteurs-individus.

En menant parallèlement ces deux analyses – analyse de N_I dans R^P, et analyse de N_J dans R^n - on aboutit à des *formules symétriques* liant les composantes principales des points-individus aux composantes principales des points-caractères.

Ces formules sont :

$$\begin{cases} f_{ri} = \dfrac{1}{\sqrt{\lambda_r}} \sum_{j=1}^{n} \dfrac{P_{ij}}{P_{i.}} g_{rj} & (i = 1, 2, \ldots, n) \\ g_{rj} = \dfrac{1}{\sqrt{\lambda_r}} \sum_{i=1}^{1} \dfrac{P_{ij}}{P_{.j}} f_{ri} & (j = 1, 2, \ldots, p) \end{cases}$$

Elles montrent que la valeur f_{ri} de la projection du i.ème point-individu sur la r.ième axe principale Δ_r (respectivement la valeur g_{rj} de la projection du j.ième point-caractère sur Δ_r) est, au facteur $\sqrt{\lambda_r}$ près, égale au barycentre des g_{rj} (respectivement des f_{ri}) affectués des masses $\dfrac{P_{ij}}{P_{i.}}$ $\left(\text{respectivement } \dfrac{P_{ij}}{P_{.j}}\right)$.

Cette propriété permet de représenter rigoureusement sur les différents plans principaux *et de façon simultanée* le nuage des points-individus et le nuage des points-caractères.

Dans cette représentation, *si les nuages ne sont pas trop déformés par projection,* deux points-individus sont d'autant plus proches que leurs profils sont semblables. Plus la distance entre deux points-caractères est petite, plus leur corrélation est grande. De plus, la proximité entre un poi t-individu et un point-caractère a un sens précis : un point-individu est proche des points-caractères avec lesquels il s'associe le plus (poids forts) et vice-versa.

Cette propriété importante *Propre à l'analyse factorielle des correspondances* permet de caractériser aisément des groupes d'individus homogènes s'ils sont mis en évidence lors de l'interprétation des résultats.

REFERENCES BIBLIOGRAPHIQUES

(Réf. 1) ANDERSON T.W. – Introduction to Multivariate statistical Analysis John Wiley, 1968

(Réf. 2) BENZECRI J.P. – Leçon sur l'analyse statistique des données multidimensionnelles
1970. Faculté des Sciences de Paris

(Réf. 2') BENZECRI J.P. – Distance distributionnelle et métrique du chi.2 en analyse factorielle des correspondances
1970. Faculté des Sciences de Paris

(Réf. 2'') BENZECRI J.P. – Analyse de tableaux binaires multiples 1971. Faculté des Sciences de Paris

(Réf. 3) CAZES P. – Etude du dédoublement d'un tableau en analyse factorielle des correspondances
1972 – Laboratoire du Professeur J.P. Benzecri. Faculté des Sciences de Paris

(Réf. 4) CAILLIEZ F., MAILLES J.P., NAKACHE J.P., PAGES J.P. – Analyse des données multidimensiennelles
1971 C.E.E.E.

(Réf. 5) DELABRE M. – Contribution à l'Etude de la Classification Automatique. Thèse pour le Doctorat en Médecine Lille 1971

(Réf. 6) KAMPE DE FERIET J. – La théorie de l'information
Publication du laboratoire de Calcul Numérique de la Faculté des Sciences de Lille 1962

(Réf. 7) LEBART L., FENELON J.P. – Statistique et Informatique Appliquée
1971 – Dunod

(Réf. 8) – MORRISSON D.F. – Multivariate Statistical Methods
1967 Mc Graw-Hill Book Compagny

(Réf. 9) NAKACHE J.P., LEBART L. – Analyse Factorielle des Correspondances
Collection M.I.S. Fasc. N° 8 1970

(Réf. 10) NAKACHE J.P. – Analyse en Composantes Principales
M.I.S. Fasc. N° 7 1970

(Réf. 11) SHANNON C.E. – A Mathematical Theory of Communications
Bell Syst. Tech. J. 1948

(Réf. 12) YAGLOM A.M., YAGLOM J.M. – Probabilité et Information (Monographies DUNOD).