L'ANALYSE DES CORRESPONDANCES MULTIPLES : PRINCIPES ET APPLICATION

R. PALM*

RÉSUMÉ

Cette note décrit les principes de l'analyse des correspondances multiples et précise les relations qui existent entre différentes analyses d'un même ensemble de données. Un exemple, traité par le logiciel SAS, illustre la méthode.

SUMMARY

This note describes the principles of multiple correspondence analysis and gives the relationship between different ways of analysing the same dataset. An example, processed by SAS software, illustrates the method.

1. INTRODUCTION

L'analyse des correspondances est, à l'origine, une méthode statistique descriptive visant à mettre en correspondance les modalités de deux variables ou caractères qualitatifs. Le point de départ de cette analyse est un tableau de contingence à deux entrées, les différentes colonnes du tableau correspondant aux p_1 modalités d'un premier caractère et les lignes du tableau correspondant aux p_2 modalités d'un deuxième caractère.

L'analyse des correspondances multiples est une extension de l'analyse des correspondances simples, qui a pour objet l'étude simultanée de plus de deux caractères. Un cas typique d'utilisation est la description de grands tableaux de données provenant, par exemple, des résultats d'enquêtes auprès de personnes. Les lignes correspondent alors aux individus enquêtés et les colonnes correspondent aux diverses questions posées.

Dans une note antérieure, l'analyse des correspondances simples a été présentée de manière assez concrète, en considérant qu'il s'agit d'une application particulière d'une méthode d'analyse plus générale [PALM, 1993].

*Professeur à la Faculté universitaire des Sciences agronomiques de Gembloux.

L'extension au cas de plus de deux variables y a également été brièvement abordée. L'objectif de la présente note est d'approfondir l'analyse des correspondances multiples, afin de donner à l'utilisateur les éléments lui permettant de procéder à l'interprétation des résultats issus d'une telle analyse.

Nous présentons d'abord la méthode dans le cas général d'un nombre quelconque de caractères et de modalités (paragraphe 2). Ensuite, nous examinons les relations existant entre différentes analyses d'un même ensemble de données dans quelques cas particuliers, liés aux nombres de caractères et aux nombres de modalités pris en compte (paragraphe 3). Nous terminons par quelques informations complémentaires (paragraphe 4) avant de tirer les conclusions (paragraphe 5).

La présentation de l'analyse des correspondances multiples se fera en supposant que le lecteur maîtrise les principes de l'analyse des correspondances simples. Dans la négative, nous renvoyons celui-ci à la note publiée antérieurement [PALM, 1993].

Par ailleurs, le lecteur désireux d'approfondir l'étude ou à la recherche d'une présentation plus formelle de l'analyse des correspondances consultera, par exemple, ESCOFIER et PAGES [1998], LEBART *et al.* [1995], SAPORTA [1990], TOMASSONE *et al.* [1993] ou VOLLE [1985].

2. ANALYSE DES CORRESPONDANCES MULTIPLES DE p CARACTÈRES

2.1. Tableau disjonctif complet et table de Burt

Soient $p_1, p_2, \ldots, p_s$ les nombres de modalités des s caractères considérés et soit p le nombre total de modalités pour les s caractères. Le tableau disjonctif complet comporte n lignes, à raison d'une ligne par individu, et p colonnes. Pour chaque ligne, la valeur 1 est notée en regard des modalités prises par l'individu et la valeur 0 est notée en regard des modalités non prises par l'individu. Chaque ligne comporte donc s valeurs unitaires et $p-s$ valeurs nulles. Pour un tel tableau, les sommes par ligne sont constantes et égales à s, les sommes des sommes des colonnes relatives aux modalités d'un caractère sont constantes et égales à n et la somme de l'ensemble des cellules du tableau est égale au produit ns.

Bien qu'on puisse, comme dans le cas de deux caractères, additionner des lignes identiques et réduire de ce fait la taille du tableau initial sans modifier l'analyse, nous considérons, dans la suite, qu'une ligne correspond à un seul individu. De même, nous ne considérons pas le cas d'un tableau incomplet, où aucune modalité n'est attribuée pour un ou plusieurs caractères pour au moins un individu. Certaines propriétés énoncées par la suite ne sont, en effet, pas vérifiées pour des tableaux incomplets.

Si $\boldsymbol{X}$ est la matrice correspondant au tableau disjonctif complet, la matrice $\boldsymbol{X}'\boldsymbol{X}$ s'appelle table de BURT. Cette table peut être partitionnée en sous-tableaux. Les sous-tableaux situés sur la diagonale principale donnent les fréquences pour les modalités d'un caractère : il s'agit de tableaux carrés dont les

éléments hors diagonale sont nuls. Quant aux sous-tableaux situés hors de la diagonale principale, ils croisent les modalités d'un caractère avec les modalités d'un autre caractère.

A titre d'illustration, nous avons soumis à l'analyse des correspondances les données publiées par HOSMER et LEMESHOW [1989] résultant d'une enquête réalisée auprès de 412 femmes, afin d'évaluer la connaissance, l'attitude et le comportement de celles-ci vis-à-vis de la mammographie. Les six caractères retenus et les modalités prises par ces caractères sont détaillés dans le tableau 1. Les données ont été traitées avec le logiciel SAS [SAS, 1989].

Tableau 1. Caractères et modalités des caractères considérés.

MAMMO : Examen mammographique
- MJAM : jamais de mammographie subie,
- M1AN : mammographie subie il y a moins d'un an,
- MP1AN : mammographie subie il y a plus d'un an.

SYMPT : Attitude vis-à-vis de l'affirmation suivante : "Une mammographie n'est pas nécessaire à moins que vous ne développiez des symptômes".
- SYMPT1 : tout à fait d'accord et d'accord,
- SYMPT2 : pas d'accord,
- SYMPT3 : absolument pas d'accord.

INTERET : Intérêt perçu de la mammographie.
- INTGRAN : intérêt grand,
- INTMOY : intérêt moyen,
- INTFAIB : intérêt faible.

ANTEC : Antécédents familiaux : mère ou soeur ayant développé un cancer du sein
- ANTOUI : oui,
- ANTNON : non.

EXASEIN : Apprentissage de l'autoexamen du sein : "Quelqu'un vous a-t-il montré comment réaliser un autoexamen des seins ?"
- EXAOUI : oui,
- EXANON : non.

PDETEC : Probabilité de détection : "Quelle est la probabilité qu'une mammographie puisse détecter un nouveau cas de cancer du sein ?"
- PDETFAIB : faible probabilité,
- PDETMOY : probabilité moyenne,
- PDETGRAN : grande probabilité.

Par rapport aux données originales, quelques modifications ont été apportées dans le codage des caractères. Ainsi, le caractère SYMPT, relatif à l'utilité de la mammographie en l'absence de symptômes est, à l'origine, un caractère à quatre modalités. Les modalités "tout à fait d'accord" et "d'accord" avec la proposition ont été fusionnées car ces deux modalités présentent des effectifs nettement plus faibles que les deux autres modalités ("pas d'accord" et "absolument pas d'accord"). Cette fusion permet de ne considérer que trois modalités, comme pour plusieurs autres caractères. D'autre part, la variable INTERET, concernant l'intérêt perçu d'une mammographie, correspond initialement à la somme des résultats de cinq questions, ces résultats étant exprimés sur une échelle à quatre points. La valeur prise par cette somme varie de 5 à 17, sur une échelle de 0 à 20, la valeur étant d'autant plus grande que l'intérêt perçu est faible. Cette échelle a été divisée en trois parties, le découpage en classes étant opéré de manière à équilibrer au mieux les effectifs des classes : les valeurs 4 et 5 constituent la modalité "intérêt grand", les valeurs 7 et 8 constituent la modalité "intérêt moyen" et les valeurs 9 à 17 correspondent à la modalité "intérêt faible". Une justification du recodage des caractères SYMPT et INTERET apparaîtra aux paragraphes 3.2 et 4.

Enfin, les codages des variables MAMMO et PDETEC ont été modifiés de manière à obtenir des résultats identiques à ceux présentés dans l'ouvrage de HOSMER et LEMESHOW [1989]. Des discordances existent en effet entre les données fournies en annexe de ce livre et la description des résultats au sein du livre.

La procédure SAS permettant de recoder les variables et de réaliser l'analyse des correspondances multiples est reprise en annexe. On notera que le tableau soumis à l'analyse n'est pas le tableau disjonctif complet. Le logiciel SAS construit en effet automatiquement le tableau disjonctif complet et la table de BURT à partir des variables reprenant les diverses modalités, grâce à l'option TABLES, associée à l'option MCA de la procédure PROC CORRESP.

La figure 1 donne la table de BURT. On constate, par exemple, que parmi les 412 personnes interrogées, 104 ont subi une mammographie depuis moins d'un an, 74 ont subi une mammographie depuis plus d'un an, et 234 n'ont jamais subi de mammographie. On constate aussi que parmi les 74 personnes ayant subi une mammographie depuis plus d'un an, 12 ont choisi la modalité SYMPT1, 32 ont choisi la modalité SYMPT2 et 30 ont choisi la modalité SYMPT3 pour le second caractère. Et ainsi de suite.

2.2. Valeurs propres et inerties

L'analyse des correspondances multiples revient à effectuer une analyse des correspondances simples du tableau disjonctif complet. Toutefois, la structure particulière de ce tableau entraîne un certain nombre de conséquences sur les résultats de l'analyse.

Ainsi, le nombre maximum de valeurs propres non nulles est égal au nombre total de modalités p dont on soustrait le nombre de caractères s.

Burt Table

	M1AN	MJAM	MP1AN	SYMPT1	SYMPT2	SYMPT3	INTFAIB	INTGRAN
M1AN	104	0	0	6	43	55	21	60
MJAM	0	234	0	94	86	54	105	74
MP1AN	0	0	74	12	32	30	20	33
SYMPT1	6	94	12	112	0	0	54	31
SYMPT2	43	86	32	0	161	0	71	53
SYMPT3	55	54	30	0	0	139	21	83
INTFAIB	21	105	20	54	71	21	146	0
INTGRAN	60	74	33	31	53	83	0	167
INTMOY	23	55	21	27	37	35	0	0
ANTNON	85	22	63	102	140	126	135	147
ANTOUI	19	1	11	10	21	13	11	20
EXANON	5	44	5	23	17	14	28	14
EXAOUI	9	19	69	89	144	125	118	153
PDETFAIB	1	1	4	10	7	1	16	1
PDETGRAN	91	14	54	55	120	114	84	138
PDETMOY	12	77	16	47	34	24	46	28

	INTMOY	ANTNON	ANTOUI	EXANON	EXAOUI	PDETFAIB	PDETGRAN	PDETMOY
M1AN	23	85	19	5	99	1	91	12
MJAM	55	220	14	44	190	13	144	77
MP1AN	21	63	11	5	69	4	54	16
SYMPT1	27	102	10	23	89	10	55	47
SYMPT2	37	140	21	17	144	7	120	34
SYMPT3	35	126	13	14	125	1	114	24
INTFAIB	0	135	11	28	118	16	84	46
INTGRAN	0	147	20	14	153	1	138	28
INTMOY	99	86	13	12	87	1	67	31
ANTNON	86	368	0	50	318	15	256	97
ANTOUI	13	0	44	4	40	3	33	8
EXANON	12	50	4	54	0	8	31	15
EXAOUI	87	318	40	0	358	10	258	90
PDETFAIB	1	15	3	8	10	18	0	0
PDETGRAN	67	256	33	31	258	0	289	0
PDETMOY	31	97	8	15	90	0	0	105

Figure 1. Table de Burt.

La somme des valeurs propres, appelée aussi inertie totale, est égale au nombre maximum de valeurs propres non nulles, divisé par le nombre de caractères et la moyenne des valeurs propres est égale à l'inverse du nombre de caractères. L'inertie totale n'est donc pas fonction des observations mais uniquement du nombre de caractères et de modalités. Il s'agit là d'une différence importante par rapport à l'analyse des correspondances simples, où l'inertie totale est fonction du degré de dépendance entre les deux caractères. Pour un tableau de contingence ordinaire, la situation d'indépendance se traduit par la nullité de l'inertie totale, alors que pour un tableau disjonctif complet, l'indépendance se traduit par l'égalité des valeurs propres et l'éloignement par rapport à la situation d'indépendance est fonction de l'importance de l'écart entre la première valeur propre et la moyenne des valeurs propres.

Les valeurs propres étant, comme en analyse des correspondances simples, toujours inférieures ou égales à l'unité, la part de l'inertie totale liée au premier axe est toujours inférieure ou égale à l'inverse de la somme des valeurs propres, l'égalité n'étant observée que si la première valeur propre est égale à l'unité. Cette part d'inertie est donc inférieure ou égale au rapport du nombre de caractères sur le nombre maximum de valeurs propres non nulles. Il en résulte que si le nombre moyen de modalités par caractère est élevé, le pourcentage d'inertie lié au premier axe, et donc aussi aux suivants, sera nécessairement faible.

Par ailleurs, l'inertie d'un caractère donné, définie comme la somme des inerties des modalités du caractère, est une fonction croissante du nombre de modalités. En effet, exprimée en pour-cent de l'inertie totale, elle est égale au nombre de modalités moins une que présente ce caractère, divisé par le nombre maximum de valeurs propres non nulles. Par conséquent, si on souhaite que les différents caractères aient des contributions à peu près identiques, il est recommandé d'avoir un nombre de modalités voisin pour chacun des caractères. Quant à l'inertie d'une modalité, exprimée en pour-cent de l'inertie totale, elle est égale au complément à l'unité de la proportion d'individus présentant la modalité, divisé par le nombre maximum de valeurs propres non nulles. Ainsi donc, pour un caractère donné, l'inertie d'une modalité est d'autant plus grande que la modalité est peu représentée. C'est pour cela qu'il faut éviter, en pratique d'inclure des modalités avec des proportions d'individus trop faibles. Il est, au contraire préférable d'utiliser des modalités présentant des effectifs à peu près constants.

Les informations données ci-dessus à propos des nombres de modalités et des poids des modalités sont particulièrement utiles lors du codage de variables continues, après répartition des individus en classes, comme nous le verrons au paragraphe 4.3.

En résumé, on retiendra donc que le nombre r de valeurs propres non nulles, l'inertie totale et l'inertie relative maximum d'un caractère ne dépendent que du nombre de caractères et du nombre total de modalités p :

$$\text{nombre de valeurs propres non nulles} = r \leq p - s \,,$$

$$\text{inertie totale} = \sum_{k=1}^{p-s} \mu_k = (p-s)/s\,,$$

$$\text{inertie relative du premier facteur} = \mu_1 \Big/ \sum_{k=1}^{p-s} \mu_k \leq s/(p-s)\,,$$

μ_k étant les valeurs propres; l'inertie relative d'un caractère q (pour p et s fixés) augmente avec le nombre p_q de modalités de ce caractère :

$$\text{inertie relative du caractère } q = (p_q - 1)/(p-s)\,;$$

et l'inertie d'une modalité augmente avec la rareté de la modalité :

$$\text{inertie de la modalité } j = \left(1 - \frac{n_{.j}}{n}\right)/(p-s)\,,$$

$n_{.j}$ étant la fréquence de la modalité j.

A l'exception de l'inertie d'une modalité, on peut constater que toutes les valeurs sont déterminées uniquement à partir de la connaissance du nombre de caractères et du nombre de modalités par caractère. Elles sont donc indépendantes des observations proprement dites et peuvent être calculées *a priori*.

Pour l'exemple considéré, on dispose de six caractères totalisant 16 modalités. Le nombre maximum de valeurs propres non nulles est par conséquent égal à 10, la somme des valeurs propres vaut $10/6$, soit 1,667, la valeur propre moyenne vaut $1/6$ soit 0,1667. L'inertie relative liée au premier facteur ne peut pas dépasser $6/10$, soit 60 %. Pour un caractère à deux modalités, l'inertie relative est égale à $1/10$ ou encore 10 % de l'inertie totale; pour un caractère à trois modalités, l'inertie est égale à $2/10$ ou 20 % de l'inertie totale.

Si on s'intéresse à l'inertie des modalités d'un caractère, il est nécessaire de connaître les fréquences relatives de ces modalités. Ainsi pour le caractère MAMMO, la table de BURT (figure 1) montre que la modalité M1AN a été sélectionnée 104 fois, la modalité MJAM a été sélectionnée 234 fois et la modalité MP1AN a été sélectionnée 74 fois. Les inerties relatives de ces trois modalités sont donc :

$$\left(1 - \frac{104}{412}\right) \Big/ 10 = 0,0748\,; \quad \left(1 - \frac{234}{412}\right) \Big/ 10 = 0,0432$$

$$\text{et} \quad \left(1 - \frac{74}{412}\right) \Big/ 10 = 0,0820\,,$$

la somme de ces trois inerties étant égale à 0,20.

On vérifie donc bien, sur cet exemple, que l'inertie d'un caractère augmente avec le nombre de modalités de ce caractère et que l'inertie d'une modalité est d'autant plus grande que la modalité est peu représentée. C'est précisément pour équilibrer les inerties liées aux modalités et aux caractères que certains codages des données originales ont été modifiés (paragraphe 2.1).

La figure 2 reprend les valeurs propres, ainsi que les informations relatives à la qualité des représentations des modalités sur le premier plan factoriel, les masses et les inerties des différentes modalités. On constate que l'inertie liée au premier facteur est de 20 % environ. On constate aussi que les inerties liées aux quatre premiers facteurs sont supérieures à l'inertie moyenne, qui est de 10 %. La première de ces valeurs propres est sensiblement plus grande que les autres, qui présentent une décroissance lente.

Les valeurs reprises dans la colonne relative à la qualité correspondent aux sommes des cosinus carrés des modalités sur les deux premiers axes, ce nombre d'axes étant le nombre retenu par défaut dans le logiciel SAS. Ces éléments seront précisés au paragraphe 2.4.

Les masses des modalités sont égales aux rapports entre les fréquences absolues de ces modalités et l'effectif total, multipliés par le nombre de caractères. Ainsi, pour les modalités de la variable MAMMO, les masses sont, pour M1AN, MJAM et MP1AN, respectivement égales à :

$$\frac{104}{(6)(412)} = 0,0421\,, \quad \frac{234}{(6)(412)} = 0,0947 \quad \text{et} \quad \frac{74}{(6)(412)} = 0,0299\,.$$

La somme des masses des modalités d'un caractère est donc toujours égale à l'inverse du nombre de caractères :

$$0,0421 + 0,0947 + 0,0299 = 1/6 = 0,1667\,.$$

En ce qui concerne les inerties des modalités et des caractères, on retrouve bien, dans la figure 2, les résultats calculés ci-dessus dans le cas du caractère MAMMO, pris à titre d'exemple.

2.3. Coordonnées des modalités sur les axes factoriels

Le centre de gravité des modalités relatives à un caractère donné se situe à l'origine des axes. Il y a donc nécessairement au moins une modalité présentant une coordonnée positive et une modalité présentant une coordonnée négative. En particulier, dans le cas d'un caractère à deux modalités, celles-ci sont situées de part et d'autre de l'origine des axes, la distance de l'origine étant inversement proportionnelle à la masse de la modalité, c'est-à-dire à la fréquence relative de la modalité, comme nous l'illustrerons au paragraphe 2.4.

Lorsqu'un caractère présente plus de deux modalités et que ces modalités peuvent être ordonnées, il peut être utile de relier les points relatifs à ces modalités par une ligne brisée, en respectant l'ordre des modalités. De telles lignes peuvent faire apparaître des relations non linéaires entre les caractères et les axes. Elles peuvent aussi mettre en évidence des comportements similaires de certains caractères.

D'autre part, la coordonnée d'une modalité d'un caractère donné est, à une constante près, la moyenne pondérée des coordonnées des individus qui présentent cette modalité, la constante en question étant l'inverse de la racine carrée

Inertia and Chi-Square Decomposition

Singular Value	Principal Inertia	Chi-Square	Percent	4 8 12 16 20
0.57273	0.32802	926.89	19.68	**********
0.44910	0.20170	569.94	12.10	**********
0.43408	0.18843	532.45	11.31	**********
0.41641	0.17339	489.96	10.40	**********
0.40567	0.16456	465.01	9.87	**********
0.39233	0.15393	434.95	9.24	**********
0.36004	0.12963	366.29	7.78	**********
0.35464	0.12577	355.38	7.55	**********
0.32356	0.10469	295.84	6.28	********
0.31073	0.09655	272.84	5.79	*******
Total	1.66667	4709.55	100.00	

Degrees of Freedom = 225

	Quality	Mass	Inertia
M1AN	0.3766	0.0421	0.0748
MJAM	0.5051	0.0947	0.0432
MP1AN	0.0551	0.0299	0.0820
SYMPT1	0.4473	0.0453	0.0728
SYMPT2	0.2732	0.0651	0.0609
SYMPT3	0.3879	0.0562	0.0663
INTFAIB	0.4980	0.0591	0.0646
INTGRAN	0.3102	0.0676	0.0595
INTMOY	0.1125	0.0400	0.0760
ANTNON	0.2014	0.1489	0.0107
ANTOUI	0.2014	0.0178	0.0893
EXANON	0.2099	0.0218	0.0869
EXAOUI	0.2099	0.1448	0.0131
PDETFAIB	0.4766	0.0073	0.0956
PDETGRAN	0.3822	0.1169	0.0299
PDETMOY	0.4181	0.0425	0.0745

Figure 2. Valeurs propres, qualité de représentation sur le premier plan factoriel, masses et inerties des modalités.

de la valeur propre associée à l'axe. Si le tableau disjonctif de départ comporte une ligne par individu, les poids sont constants et égaux à $1/n_{.j}$. Si les individus identiques ont été regroupés, les poids sont égaux aux fréquences relatives des catégories d'individus qui présentent la modalité.

De manière similaire, la coordonnée d'un individu sur un axe est, à une constante près la moyenne pondérée des coordonnées des modalités qui le caractérisent. La constante est l'inverse de la racine carrée de la valeur propre associée à l'axe et les poids sont constants et égaux à l'inverse du nombre de caractères, puisque, pour un tableau disjonctif complet, le total d'une ligne est égal au nombre de caractères.

Le fait que, à une constante près, un point-colonne (ou un point-ligne) soit le barycentre des points-lignes (ou des points-colonnes) n'est pas spécifique à l'analyse d'un tableau disjonctif complet. Mais dans ce dernier cas, les profils des lignes et des colonnes contiennent de nombreuses valeurs nulles. Il en résulte que les coordonnées des individus qui ne présentent pas une modalité donnée n'interviennent pas dans le calcul de la coordonnée de cette modalité, leurs poids étant nuls. De même, les coordonnées des modalités non observées sur un individu n'interviennent pas dans le calcul de la coordonnée de cet individu, leurs poids étant nuls.

La figure 3 donne les coordonnées sur les deux premiers axes factoriels des 16 modalités de l'exemple considéré et la figure 4 donne la représentation de ces modalités dans le premier plan factoriel. L'interprétation des résultats sera réalisée au paragraphe 2.5.

On peut noter que les figures 3 et 4 ne donnent aucune information concernant les individus, c'est-à-dire les lignes du tableau disjonctif complet.

Bien que la représentation simultanée des individus et des modalités puisse présenter un intérêt pour l'interprétation des résultats, elle n'est en général pas réalisée, car les individus sont souvent très nombreux. Même s'il existe des groupes d'individus de profil identique conduisant à des points multiples, le nombre de profils différents risque d'être important lorsque les modalités sont nombreuses. Il en résulterait des graphiques peu clairs et des documents de résultats fort volumineux.

Dans la plupart des applications, les individus sont d'ailleurs anonymes et ne présentent pas d'intérêt par eux-mêmes. Ainsi, la représentation des 412 femmes ayant participé à l'étude utilisée comme illustration est sans intérêt direct, dans la mesure où elles constituent un échantillon d'une population plus vaste.

Par contre, la représentation de quelques individus typiques, sous la forme d'individus supplémentaires peut se faire sans difficulté, notamment dans le but de faciliter l'interprétation, comme nous le signalerons au paragraphe 4.1.

A titre d'exemple, considérons une personne caractérisée par les modalités suivantes :

M1AN, SYMPT3, INTGRAN, ANTOUI, EXAOUI et PDETGRAN.

Column Coordinates

	Dim1	Dim2
M1AN	-1.0150	0.2914
MJAM	0.5774	-0.2255
MP1AN	-0.3993	0.3035
SYMPT1	0.9958	-0.4544
SYMPT2	-0.0183	0.6523
SYMPT3	-0.7812	-0.3895
INTFAIB	0.7983	0.5197
INTGRAN	-0.6668	-0.1021
INTMOY	-0.0524	-0.5942
ANTNON	0.0695	-0.1387
ANTOUI	-0.5816	1.1603
EXANON	1.1003	0.4252
EXAOUI	-0.1660	-0.0641
PDETFAIB	1.8614	2.6395
PDETGRAN	-0.3783	0.1398
PDETMOY	0.7222	-0.8372

Figure 3. Coordonnées des modalités sur les deux premiers axes factoriels.

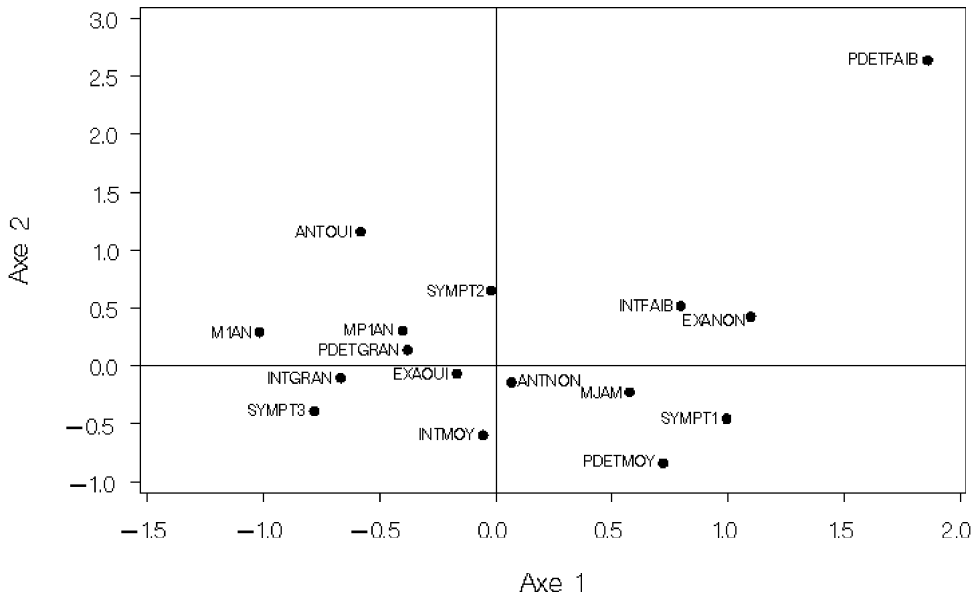

Figure 4. Représentation des modalités dans le premier plan factoriel.

La coordonnée de ce point-ligne sur le premier axe serait :

$$(-1,0150-0,7812-0,6668-0,5816-0,1660-0,3783)/(6\sqrt{0,32802}) = -1,0444\,.$$

Sur la figure 3, cet individu se situerait donc, pour le premier facteur, légèrement plus à gauche que la modalité M1AN. Nous rediscuterons de la position de cet individu, lors de l'interprétation des facteurs (paragraphe 2.5).

2.4. Contribution aux facteurs et cosinus carrés

Comme pour l'analyse des correspondances simples, on peut calculer la contribution relative d'un point à un facteur, en multipliant le carré de la coordonnée du point sur l'axe par la masse et en divisant le résultat par l'inertie de l'axe. Cette contribution relative quantifie l'importance relative d'un point dans la définition du facteur et, lors de l'interprétation des résultats, on repère d'abord les modalités qui ont une forte contribution aux axes.

Dans le cas de l'analyse des correspondances multiples, en plus des contributions relatives des différentes modalités, on peut calculer la contribution de chacun des caractères au facteur, en additionnant les contributions des diverses modalités d'un caractère.

Ainsi, pour la modalité M1AN, la contribution au premier facteur est égale à :

$$(-1,0150^2)\,(0,0421)/0,32802 = 0,1321\,.$$

De la même manière, on trouve que les contributions des modalités MP1AN et MJAM au premier facteur sont égales à 0,0146 et 0,0962 et que, par conséquent la contribution du caractère MAMMO au facteur est égale à :

$$0,1321 + 0,0146 + 0,0962 = 0,2429 \;\text{ soit }\; 24\,\%\,.$$

La figure 5 donne les contributions relatives des modalités aux deux premiers axes. Pour les modalités du caractère MAMMO, on retrouve bien les valeurs calculées ci-dessus.

Quant aux cosinus carrés, ils quantifient la qualité de la représentation d'un point-ligne ou d'un point-colonne sur un axe factoriel, sur un plan factoriel ou, de façon plus générale, dans un sous-espace de l'espace factoriel complet. Ces cosinus carrés s'obtiennent de la même manière qu'en analyse des correspondantes simples. Soit z_{jk} la coordonnée de la modalité j ($j = 1, \ldots, p$) sur l'axe k, le cosinus carré pour cet axe s'écrit :

$$\cos^2 = z_{jk}^2 \Big/ \sum_{k=1}^{r} z_{jk}^2\,.$$

L'addition des cosinus carrés relatifs aux facteurs successifs donne la qualité de la représentation des modalités dans des espaces de dimension croissante.

Partial Contributions to Inertia for the Column Points

	Dim1	Dim2
M1AN	0.1321	0.0177
MJAM	0.0962	0.0239
MP1AN	0.0146	0.0137
SYMPT1	0.1370	0.0464
SYMPT2	0.0001	0.1374
SYMPT3	0.1046	0.0423
INTFAIB	0.1147	0.0791
INTGRAN	0.0916	0.0035
INTMOY	0.0003	0.0701
ANTNON	0.0022	0.0142
ANTOUI	0.0184	0.1188
EXANON	0.0806	0.0196
EXAOUI	0.0122	0.0030
PDETFAIB	0.0769	0.2515
PDETGRAN	0.0510	0.0113
PDETMOY	0.0675	0.1476

Squared Cosines for the Column Points

	Dim1	Dim2
M1AN	0.3479	0.0287
MJAM	0.4383	0.0668
MP1AN	0.0349	0.0202
SYMPT1	0.3702	0.0771
SYMPT2	0.0002	0.2730
SYMPT3	0.3107	0.0772
INTFAIB	0.3498	0.1482
INTGRAN	0.3031	0.0071
INTMOY	0.0009	0.1117
ANTNON	0.0404	0.1610
ANTOUI	0.0404	0.1610
EXANON	0.1826	0.0273
EXAOUI	0.1826	0.0273
PDETFAIB	0.1583	0.3183
PDETGRAN	0.3363	0.0459
PDETMOY	0.1784	0.2397

Figure 5. Contributions relatives et cosinus carrés des modalités pour les deux premiers axes factoriels.

Les cosinus carrés des modalités sont repris dans la figure 5. Si on additionne les cosinus carrés d'une modalité, pour les deux facteurs retenus, on obtient les valeurs reprises sous l'intitulé "Quality" de la figure 2.

On constate qu'aucune modalité n'est vraiment bien représentée sur le premier axe factoriel, la valeur la plus élevée étant de 0,44 pour la modalité MJAM. La prise en compte du deuxième axe permet essentiellement d'augmenter la qualité de la représentation des modalités les plus mal représentées sur le premier axe. Comme l'a montré le tableau des valeurs propres (figure 2), même en considérant les quatre premiers axes, la qualité des représentations reste, en moyenne, assez faible, puisque la somme des inerties liées aux quatre facteurs n'est que de 53 %.

Dans le cas d'un caractère ne présentant que deux modalités, les valeurs de la coordonnée, de la contribution et du cosinus carré d'une modalité sont directement liées aux valeurs correspondantes de l'autre modalité. En effet, soit $f_{.1}$ et $f_{.2}$ les masses des deux modalités et z_{1k} et z_{2k} les coordonnées des deux modalités sur l'axe k. Le centre de gravité des deux modalités étant nul, on a :

$$|f_{.1}\, z_{1k}| = |f_{.2}\, z_{2k}|\,,$$

ou encore :

$$\frac{z_{1k}}{z_{2k}} = \frac{f_{.2}}{f_{.1}}\,.$$

Les contributions à l'axe k sont égales à :

$$f_{.1} z_{1k}^2/\mu_k \quad \text{et} \quad f_{.2} z_{2k}^2/\mu_k\,,$$

soit un rapport égal à $f_{.2}/f_{.1}$. Les cosinus carrés valent :

$$z_{1k}^2 \Big/ \sum_{k=1}^{r} z_{1k}^2 \quad \text{et} \quad z_{2k}^2 \Big/ \sum_{k=1}^{r} z_{2k}^2\,.$$

En multipliant le numérateur et le dénominateur des deux quantités respectives par $f_{.1}^2$ et $f_{.2}^2$, on constate que ces deux valeurs sont toujours identiques, puisque $|f_{.1} z_{1k}| = |f_{.2} z_{2k}|$, comme nous venons de le voir ci-dessus.

En se rappelant que les masses des modalités sont proportionnelles aux fréquences de ces modalités, on peut donc affirmer que, dans le cas de caractères à deux modalités, les coordonnées des modalités sur un axe sont nécessairement de signe opposé et que l'éloignement de l'origine des modalités est inversement proportionnel à la fréquence relative de la modalité, que les contributions relatives de ces deux modalités sont également inversement proportionnelles à leur fréquence mais que les cosinus carrés sont égaux.

A titre d'exemple, considérons les deux modalités EXAOUI et EXANON de la variable EXASEIN. La table de BURT (figure 1) nous montre que la modalité EXAOUI a été retenue 358 fois et que la modalité EXANON a été retenue 54 fois. Le rapport de la deuxième fréquence à la première fréquence est égal à 0,151.

Les coordonnées sur le premier axe sont égales à $-0,1660$ et 1,1003 (figure 3) et le rapport est bien égal, en valeur absolue, à 0,151. Pour les contributions des modalités à cet axe, on a respectivement 0,0122 et 0,0806 et le rapport est, ici aussi, égal à 0,151. Enfin, les cosinus carrés sont tous deux égaux à 0,1826 (figure 5).

2.5. Interprétation des facteurs

L'interprétation des coordonnées, des contributions et des cosinus carrés des points-lignes ou colonnes est sensiblement la même qu'en analyse des correspondances simples.

En pratique, on repère les modalités et les caractères qui contribuent le plus aux axes factoriels. De même, si on dispose des informations relatives aux individus, on peut repérer ceux qui ont les contributions les plus importantes aux différents facteurs.

La proximité des modalités de caractères différents s'interprète en termes d'association, deux modalités étant d'autant plus proches qu'elles ont été retenues, globalement, par les mêmes individus. Deux modalités qui sont choisies par les mêmes individus coïncident.

La proximité de deux modalités d'un même caractère traduit une ressemblance de ces deux modalités. Pour un tableau disjonctif complet, deux modalités ne peuvent pas être sélectionnées par les mêmes individus et le centre de gravité des modalités d'un caractère donné est situé à l'origine des axes. Deux modalités proches ont donc nécessairement été choisies par deux groupes d'individus différents, mais ces deux groupes présentent des ressemblances vis-à-vis d'autres caractères.

Enfin, la proximité de deux individus s'interprète également en termes de ressemblance dans le choix des modalités retenues par chacun d'eux. Deux individus ayant retenu exactement les mêmes modalités sont, en effet, confondus dans les représentations graphiques.

Dans le cas de l'exemple considéré, l'addition des contributions relatives des modalités d'un même caractère, fait apparaître que quatre caractères ont une contribution comprise, pour le premier facteur, entre 20 et 25 %, les caractères EXASEIN et ANTEC ayant respectivement une contribution de 9 et de 2 %.

Si on note pour chaque caractère la modalité ayant la coordonnée négative la plus grande en valeur absolue, sur le premier axe, on obtient la liste suivante :

M1AN, SYMPT3, INTGRAN, ANTOUI, EXAOUI et PDETGRAN.

A l'inverse, les modalités ayant la coordonnée positive la plus grande sont :

MJAM, SYMPT1, INTFAIB, ANTNON, EXANON et PDETFAIB.

Pour les caractères à trois modalités, les modalités intermédiaires MP1AN, INTMOY, PDETMOY et SYMPT2 occupent des positions intermédiaires. Ces

modalités intermédiaires ont des contributions partielles moins importantes au premier facteur que les modalités extrêmes, sauf PDETMOY, qui a une contribution un peu supérieure à la modalité PDETGRAN.

Si on exclut de l'interprétation la variable ANTEC, dont la contribution au premier facteur n'est que de 2 %, on peut interpréter ce facteur comme un gradient de sensibilisation des femmes vis-à-vis de l'aspect préventif de la mammographie, l'axe étant orienté, de façon arbitraire, d'une forte sensibilisation (valeurs négatives sur l'axe) vers une faible sensibilisation (valeurs positives sur l'axe).

Pour confirmer cette interprétation, on peut calculer les coordonnées de deux individus supplémentaires particuliers. Le premier individu est caractérisé par les modalités suivantes : M1AN, SYMPT3, INTGRAN, ANTOUI, EXAOUI et PDETGRAN. Il s'agit de l'individu supplémentaire dont la coordonnée, qui a été déterminée au paragraphe 2.3, est égale à $-1,0444$. Il correspond à une femme ayant réalisé une mammographie il y a moins d'un an, estimant qu'il n'est pas nécessaire de développer des symptômes pour réaliser une mammographie, ayant des antécédents familiaux, ayant reçu un apprentissage de l'autoexamen du sein et considérant que la probabilité de détecter un nouveau cancer du sein par mammographie est grande. Il s'agit donc bien d'une femme caractérisée par une grande sensibilisation vis-à-vis de l'aspect préventif de la mammographie.

Le deuxième individu est caractérisé par les modalités MJAM, SYMPT1, INTFAIB, ANTNON, EXANON et PDETFAIB. Pour cette personne, la coordonnée sur le premier axe factoriel, qui se calcule comme expliqué au paragraphe 2.3, vaut 1,5722. Cette personne est peu sensibilisée à l'aspect préventif de la mammographie.

L'analyse de la table de Burt confirme l'interprétation qui vient d'être donnée pour le premier facteur. Si on examine, par exemple, la table croisant les modalités de la variable MAMMO et de la variable SYMPT, on constate que, parmi les femmes ayant réalisé une mammographie depuis moins d'un an, 52 % présentent la modalité SYMPT3, alors qu'elles ne sont que 23 % parmi celles qui n'ont jamais réalisé de mammographie. Inversement, elles ne sont que 6 % à présenter la modalité SYMPT1 si elles ont subi la mammographie depuis moins d'un an, alors qu'elles sont 16 % à présenter cette modalité si elles n'ont jamais subi de mammographie. De la même manière, on pourrait examiner plus en détail les croisements des autres caractères.

Sur le deuxième axe, les coordonnées positives les plus grandes concernent les modalités PDETFAIB et ANTOUI, et la coordonnée négative la plus importante s'observe pour la modalité PDETMOY. Par ailleurs, on constate aussi que le deuxième axe oppose les modalités extrêmes à la modalité intermédiaire pour les caractères SYMPT, INTERET et PEDETEC. Toutefois, les modalités intermédiaires ont une coordonnée négative pour les deux derniers caractères, alors qu'elle est positive pour le premier caractère.

Les variables ayant les contributions les plus importantes au deuxième facteur sont SYMPT (23 %) et PDETEC (41 %). Ces contributions sont principalement dues aux modalités SYMPT2 (14 %), PDETFAIB (25 %) et PDETMOY (15 %).

A ces modalités présentant une contribution importante au deuxième facteur, on peut encore ajouter la modalité ANTOUI (12 %).

Ce deuxième facteur oppose les modalités PDETFAIB, SYMPT2 et ANTOUI à la modalité PDETMOY. Il résulte principalement du fait que la modalité PDETFAIB est retenue par une proportion plus grande des personnes parmi celles présentant des antécédents familiaux (6,8 %) que parmi celles sans antécédents (4,0 %), alors que pour la modalité PDETMOY, on observe la situation inverse (18,2 % des personnes avec antécédents et 26,4 % des personnes sans antécédents). Il est dû également à la plus grande proportion de personnes retenant la modalité SYMPT2 parmi les personnes sans antécédents (47,7 %) que parmi les personnes avec antécédents (38,0 %).

Pour confirmer cette interprétation du deuxième facteur, on peut calculer la coordonnée sur cet axe de deux individus particuliers : le premier individu serait caractérisé par les modalités ANTOUI, SYMPT2 et PDETFAIB, alors que le deuxième individu aurait les modalités ANTNON, SYMPT1 et PDETMOY. Pour les trois autres caractères, on considère que les modalités seraient identiques pour les deux individus, la modalité retenue étant celle qui présente la coordonnée sur le deuxième axe qui est la plus faible en valeur absolue : MJAM, EXAOUI, INTGRAN.

Pour ces deux individus, les coordonnées seraient respectivement égales à $1,5068$ et $-0,6762$. Le premier individu supplémentaire se situerait en haut du graphique et le deuxième se situerait vers le bas.

Au delà de ces constatations, il semble bien difficile de donner une signification plus globale à ce facteur, dont l'inertie relative est de 12 %, soit une valeur juste supérieure à l'inertie moyenne, qui est de 10 %.

3. EQUIVALENCE DE DIFFÉRENTES ANALYSES

3.1. Analyse des correspondances simples

Le point de départ naturel de l'analyse des correspondances simples est un tableau de contingence à deux entrées, les p_1 modalités d'un premier caractère étant croisées avec les p_2 modalités d'un second caractère.

L'analyse peut cependant aussi être menée à partir du tableau disjonctif complet comportant n lignes et $p = p_1 + p_2$ colonnes. La comparaison des résultats obtenus, d'une part, par l'analyse du tableau de contingence et, d'autre part, par l'analyse du tableau disjonctif complet a été présentée antérieurement [PALM, 1993]. Nous rappelons simplement les éléments les plus importants de cette comparaison.

L'analyse du tableau de contingence conduit à un nombre maximum de valeurs propres non nulles égales à la plus petite dimension du tableau, diminué d'une unité. D'autre part, la somme des valeurs propres multipliée par l'effectif total n est égale à la valeur χ^2_{obs} relative au test d'indépendance des deux caractères et chaque valeur propre, exprimée en proportion de la somme des valeurs

propres, est égale à la contribution au χ^2_{obs} de chacun des facteurs.

Dans le cas de l'analyse du tableau disjonctif complet, le nombre maximum de valeurs propres non nulles est égal à $p-2$ et la somme de valeurs propres est égale à $(p-2)/2$ (paragraphe 2.2). Par conséquent, la moyenne des valeurs propres est égale à 0,5. Pour ce type d'analyse, le pourcentage d'inertie liée à un facteur, qui est souvent très faible, n'a pas d'intérêt pour l'interprétation et on ignore généralement les facteurs correspondant aux valeurs propres inférieures à 0,5, comme nous le verrons au paragraphe 4.2.

En désignant par μ_k les valeurs propres résultant de l'analyse du tableau disjonctif complet et par λ_k les valeurs propres obtenues par l'analyse du tableau de contingence, on a, pour les valeurs μ_k supérieures à 0,5 :

$$\lambda_k = (2\mu_k - 1)^2 \,.$$

Les représentations graphiques des modalités des deux caractères (points-lignes et points-colonnes pour le tableau de contingence; points-colonnes pour le tableau disjonctif complet) présentent de nettes similitudes pour les deux analyses. En effet, par rapport au cas du tableau disjonctif complet, les coordonnées sur un axe pour le tableau de contingence sont multipliées par un facteur constant, égal à la racine carrée du rapport de λ_k à μ_k.

Pour un tableau de contingence, les individus ne font pas l'objet d'une représentation graphique. Par contre, dans le cas de l'analyse à partir du tableau disjonctif complet, la coordonnée d'un individu sur l'axe factoriel k est, au facteur $1/(2\sqrt{\mu_k})$ près, égale à la moyenne des coordonnées des deux modalités présentes sur l'individu.

Pour le tableau de contingence, la somme des contributions relatives des modalités d'un caractère (somme des contributions relatives des lignes ou des colonnes) est égale à l'unité. Pour le tableau disjonctif complet, la somme des contributions relatives des modalités d'un caractère est égale à 0,5. La contribution d'une modalité est donc deux fois plus forte pour le tableau de contingence.

A titre d'illustration, nous avons repris l'exemple examiné précédemment, en ne considérant que les variables MAMMO et SYMPT, qui ont chacune trois modalités. L'analyse a été réalisée, d'une part, à partir du tableau disjonctif complet et, d'autre part, à partir du tableau de contingence. Les résultats de ces deux analyse ne sont pas reproduits ici de manière exhaustive. Nous reprenons ci-après uniquement quelques éléments, retenus comme exemples. L'analyse des correspondances du tableau disjonctif complet conduit aux quatre valeurs propres suivantes :

$$\mu_1 = 0,6844 \quad \mu_2 = 0,5130 \quad \mu_3 = 0,4870 \quad \text{et} \quad \mu_4 = 0,3156 \,,$$

dont la somme est égale à 2. L'analyse du tableau de contingence de ces mêmes données conduit aux deux valeurs propres suivantes :

$$\lambda_1 = [(2)(0,6844) - 1]^2 = 0,1360 \quad \text{et} \quad \lambda_2 = [(2)(0,5130) - 1]^2 = 0,0007 \,.$$

Les pourcentages d'inertie liée au premier axe, pour ces deux analyses sont respectivement égaux à :

$$0,6844/2 = 0,3422 \quad \text{et} \quad (0,1360)/(0,1360 + 0,0007) = 0,995\,.$$

Alors que pour l'analyse du tableau de contingence ce pourcentage possède une interprétation (contribution du facteur au χ^2_{obs}), il n'a pas de signification lorsque l'analyse est réalisée directement à partir du tableau disjonctif complet. Nous reviendrons sur ce problème au paragraphe 4.2.

Pour le tableau disjonctif complet, la coordonnée de la modalité M1AN sur le premier axe est égale à 1,1828 et sa contribution à l'axe est de 0,2580. Pour le tableau de contingence, ces valeurs deviennent :

$$(1,1828)\sqrt{0,1360/0,6844} = 0,5273 \quad \text{et} \quad (0,2580)(2) = 0,5160\,.$$

3.2. Analyse des correspondances de la table de Burt

Au paragraphe 3.1, nous avons discuté de la similitude qui existe entre l'analyse des correspondances réalisée sur le tableau de contingence et sur le tableau disjonctif complet, dans le cas de deux caractères.

Pour un nombre quelconque de caractères, il existe également des similitudes entre l'analyse réalisée à partir du tableau disjonctif complet et l'analyse réalisée à partir de la table de Burt. Ainsi, les valeurs propres ν_k obtenues à partir de la table de Burt sont égales aux carrés de valeurs propres μ_k obtenues à partir du tableau disjonctif complet :

$$\nu_k = \mu_k^2\,.$$

Les coordonnées des modalités pour l'analyse de la table de Burt sont égales aux coordonnées des modalités pour l'analyse à partir du tableau disjonctif complet, multipliées par la racine carré de la valeur propre relative au facteur pour l'analyse du tableau disjonctif complet. Les contributions sont, par contre, identiques pour les deux analyses.

L'analyse des correspondances simples de la table de Burt pour l'exemple du paragraphe 3.1 conduirait à une première valeur propre égale à :

$$\nu_1 = (0,6844^2) = 0,4684\,,$$

la coordonnée de la modalité M1AN serait égale à :

$$1,1828\sqrt{0,6844} = 0,9785\,,$$

Des calculs identiques pourraient être réalisés pour les autres valeurs propres et les autres modalités et la contribution de la modalité au premier axe serait, pour les deux analyses, égale à 0,2580.

L'analogie entre les deux analyses, qui vient d'être illustrée dans le cas de deux caractères se vérifie pour un nombre quelconque de caractères.

3.3. Cas de deux modalités par caractère

Dans certaines applications, tous les caractères pris en compte dans l'analyse ne présentent que deux modalités. C'est le cas, par exemple, lors de l'analyse d'un tableau de présence/absence de plantes dans des études phytosociologiques, ou de réponses de type oui/non lors du dépouillement d'enquêtes sociologiques.

L'analyse des correspondances multiples de ces données donne lieu à des résultats présentant quelques particularités, qui se déduisent facilement de la situation générale, en notant que :

$$p_1 = p_2 = \ldots = p_s = 2 \quad \text{et} \quad p = 2s\,,$$

p étant le nombre total de modalités et s le nombre de caractères.

Ainsi, le nombre maximum de valeurs propres non nulles est égal au nombre de caractères et la somme des valeurs propres est égale à l'unité. La moyenne des valeurs propres est par conséquent égale à l'inverse du nombre de caractères.

D'autre part, la discussion concernant les coordonnées, les contributions et les cosinus carrés des modalités d'un caractère à deux modalités qui a été présentée au paragraphe 2.4 s'applique, dans le cas présent, à l'ensemble des caractères.

On peut montrer également que, lorsque tous les s caractères ont deux modalités, l'analyse des correspondances multiples se ramène à une analyse en composantes principales sur s variables 0/1, chacune de ces variables correspondant à la présence/absence d'une modalité particulière du caractère. Dans ce cas, les valeurs propres de la matrice de corrélation seront égale aux valeurs propres issues de l'analyse des correspondances du tableau disjonctif complet, multipliées par le nombre de caractères s. Les pourcentages d'inertie associée à chaque axe sont par conséquent identiques dans les deux analyses. Les coordonnées des individus sur un axe obtenu par l'analyse en composantes sont, après division par $\sqrt{s}$, identiques aux coordonnées des lignes sur le même axe à l'issue de l'analyse des correspondances.

Par ailleurs, il existe aussi des relations entre les corrélations des variables initiales et des scores en analyse en composantes principales et les caractéristiques des modalités de l'analyse des correspondances. En effet, le carré du coefficient de corrélation d'une variable avec un axe donné de l'analyse en composantes est identique aux cosinus carré de chacune des modalités du même caractère sur l'axe provenant de l'analyse des correspondances et le carré du coefficient de corrélation d'une variable avec un axe de l'analyse en composantes, divisé par la valeur propre relative à cet axe est identique à la contribution relative du caractère pour le même axe de l'analyse des correspondances.

4. INFORMATIONS COMPLÉMENTAIRES

4.1. Individus et variables supplémentaires

Nous avons signalé, au paragraphe 2.3, que la coordonnée d'un individu sur un axe est égal à la moyenne arithmétique des coordonnées sur cet axe des

modalités prises par l'individu, divisée par le produit de la racine carrée de la valeur propre associée à l'axe et du nombre de caractères. A cette occasion, nous avons vu que la prise en considération d'individus typiques (points-lignes supplémentaires) peut apporter une aide à l'interprétation des facteurs. Ces individus typiques sont particulièrement utiles lorsque, comme c'est souvent le cas, on ne dispose pas des informations relatives aux lignes du tableau de données (paragraphe 2.3).

La notion de point supplémentaire s'applique également à des variables. Les variables supplémentaires sont des variables qu'on souhaite représenter sur les plans factoriels, mais qui n'ont pas participé à la construction de ceux-ci. Elles sont parfois dénommées variables passives, par opposition aux variables actives, qui ont participé aux calculs de valeurs et de vecteurs propres.

Le recours à des variables passives se justifie, par exemple, lorsqu'on souhaite expliquer certaines variables par d'autres, les variables à expliquer étant les variables supplémentaires. La prise en compte de variables supplémentaires peut se justifier aussi dans le but de conforter l'interprétation des axes au moyen d'un information externe. Enfin, pour des variables quantitatives, la prise en compte sous la forme de variables supplémentaires est la seule solution, lorsqu'on ne souhaite pas rendre ces dernières qualitatives par le découpage en classes.

La représentation sur un axe factoriel d'une modalité d'une variable qualitative supplémentaire repose sur la relation barycentrique vue précédemment (paragraphe 2.3). La coordonnée sur un axe donné de la modalité j de la variable est, à une constante près, égale à la moyenne des coordonnées sur cet axe des $n_{.j}$ individus parmi les n qui présentent cette modalité, la constante étant l'inverse de la racine carrée de la valeur propre relative à l'axe.

En relation avec la représentation des diverses modalités d'une variable qualitative supplémentaire, les deux questions suivantes peuvent se poser en pratique :

- existe-t-il une liaison significative entre la variable qualitative et un axe factoriel particulier et
- peut-on considérer qu'une modalité particulière d'une variable qualitative est significativement différente de la moyenne générale sur un axe donné ?

La réponse à la première question est donnée par une analyse de la variance des coordonnées des n individus sur l'axe, le critère de classification étant les modalités de la variable supplémentaire. Le test ne sera cependant correct que si les conditions d'application de l'analyse de la variance à un critère sont remplies : il faut que l'échantillon soit aléatoire et simple et que les variances résiduelles soient égales.

Pour vérifier si une modalité particulière est significativement différente de la moyenne générale, LEBART *et al.* [1995] proposent de calculer une valeur-test, t_{jk}, égale à la coordonnée de la modalité j sur l'axe k, multipliée par le facteur :

$$\sqrt{\frac{n_j(n-1)}{n-n_j}},$$

n_j étant le nombre d'individus, parmi les n, qui présentent la modalité j. Si n_j est assez grand, on peut considérer que t_{jk} est une valeur observée d'une variable normale réduite. On considère dès lors comme significativement différentes de la moyenne générale, au niveau $\alpha = 0,05$, les modalités pour lesquelles la valeur-test est supérieure, en valeur absolue, à 2.

L'origine de cette valeur-test peut se justifier de la manière suivante. Les coordonnées des n individus sur un axe donné k sont de moyenne nulle et de variance égale à la valeur propre associée à l'axe, μ_k. Si l'hypothèse nulle est vraie, c'est-à-dire si la modalité j d'une variable supplémentaire n'est pas significativement différente de la moyenne générale, les n_j individus qui présentent cette modalité constituent alors un échantillon aléatoire de n_j individus prélevés parmi n sans remise. La moyenne et la variance de la moyenne de cet échantillon sont respectivement égales à zéro et à :

$$\frac{n - n_j}{n - 1} \frac{\mu_k}{n_j} .$$

La coordonnée de la modalité j étant égale à la moyenne des coordonnées des n_j individus, divisée par la racine carrée de la valeur propre, elle correspondra donc à la réalisation d'une variable de moyenne nulle et d'écart-type égal à :

$$\sqrt{\frac{n - n_j}{(n-1)n_j}} .$$

De plus, en vertu de la normalité asymptotique de la distribution d'échantillonnage de la moyenne, on peut dire que cette variable est approximativement normale. Par conséquent, sous l'hypothèse nulle, la valeur-test obtenue en multipliant la coordonnée de la modalité par $\sqrt{n_j(n-1)/(n-n_j)}$ est une valeur observée d'une variable normale réduite.

Il en résulte aussi que l'utilisation des valeurs-tests n'a de sens que pour des modalités de variables supplémentaires. En effet, pour une modalité active, on ne peut pas considérer que les n_j individus sont pris au hasard parmi les n individus, puisque la modalité contribue à la construction de l'axe.

Pour illustrer cette notion de variable qualitative supplémentaire, nous reprenons les résultats de l'analyse des correspondances donnés au paragraphe 3.1. Cette analyse portait sur deux variables à trois modalités. Nous considérons une troisième variable, la variable INTERET, à trois modalités comme variable supplémentaire.

La coordonnée moyenne sur le premier axe des 146 individus correspondant à la modalité INTFAIB est égale à $0,3220$. La coordonnée de la modalité sur le premier axe est donc égale à :

$$(0,3220)/\sqrt{0,6844} = 0,3892 .$$

Pour les modalités INTMOY et INTGRAN, les moyennes sont égales à $-0,0078$ et $-0,2768$. Les coordonnées sur le premier axe valent donc :

$$-0,0078/\sqrt{0,6844} = -0,0095 \quad \text{et} \quad -0,2768/\sqrt{0,6844} = -0,3346 ,$$

De manière analogue, on pourrait calculer les coordonnées de ces modalités sur les autres axes.

L'analyse de la variance des coordonnées sur l'axe 1 des individus réalisée en prenant la variable INTERET comme critère de classification conduit à une valeur F_{obs} de 22,49, soit une valeur très hautement significative. On peut donc conclure, sur la base de cette analyse, qu'il existe une relation significative entre la variable INTERET et l'axe factoriel : les femmes qui considèrent que l'intérêt de la mammographie est faible se situent en moyenne plus à droite (moyenne égale à 0,32), celles qui considèrent que l'intérêt est grand se situent plus à gauche (moyenne égale à $-0,28$) et enfin, celles qui considèrent que l'intérêt est limité occupent une position intermédiaire (moyenne proche de 0).

Les valeurs-tests pour les trois modalités sont égales à :

$$(0,3892)\sqrt{\frac{146(412-1)}{412-146}} = 5,85, \quad (-0,0095)\sqrt{\frac{99(412-1)}{412-99}} = -0,11$$

et

$$(-0,3346)\sqrt{\frac{167(412-1)}{412-167}} = -5,60\,.$$

Ces valeurs confirment bien que les deux modalités extrêmes diffèrent de la moyenne.

Lorsque la variable supplémentaire est une variable quantitative, on calcule la corrélation de cette variable avec les coordonnées des individus sur les axes factoriels. Le point peut alors être positionné dans les plans factoriels, comme en analyse en composantes principales. Le carré de la distance du point au centre de gravité est une mesure de la qualité de la représentation du point dans le plan considéré et la position du point dans un plan s'interprète en terme de direction. LEBART *et al.* [1995] considèrent cependant que le découpage en classes d'une variable continue et son traitement ultérieur comme une variable nominale apporte souvent plus d'informations que la seule position de la variable continue, notamment pour la détection d'éventuelles liaisons non linéaires.

4.2. Nombre d'axes factoriels à retenir

Rappelons tout d'abord que l'inertie totale en analyse des correspondances multiples n'a pas de signification statistique, contrairement à l'inertie calculée en analyse des correspondances d'un tableau de contingence. Nous avons signalé également que les inerties liées aux différents facteurs sont en général assez faibles (paragraphe 2.2). En particulier, dans le cas de deux caractères, les pourcentages d'inertie obtenus à partir du tableau disjonctif complet sont beaucoup plus réduits que les pourcentages obtenus pour les mêmes facteurs lors de l'analyse du tableau de contingence (paragraphe 3.2).

Les taux d'inertie liés à un tableau disjonctif complet donnent une idée pessimiste de la part d'information représentée et BENZÉCRI [1979] a proposé la

formule suivante pour permettre une meilleure appréciation des taux d'inertie :

$$\text{intertie corrigée} = \left(\frac{s}{s-1}\right)^2 \left(\mu_k - \frac{1}{s}\right)^2 \quad \text{pour} \quad \mu_k > \frac{1}{s}\,,$$

s étant le nombre de caractères et μ_k les valeurs propres obtenues par l'analyse du tableau disjonctif complet. Lorsqu'on ne dispose que de deux caractères, la formule ci-dessus redonne les inerties λ_k obtenues par l'analyse du tableau de contingence (paragraphe 3.1).

L'application de cette relation aux inerties reprises dans la figure 2 donne pour les valeurs supérieures à $1/6$, les résultats suivants :

$$0,03749\,, \quad 0,00177\,, \quad 0,00068 \quad \text{et} \quad 0,00007\,.$$

Le premier axe représente par conséquent 93,7 % de l'information, le deuxième axe 4,4 %, le troisième axe 1,7 % et le quatrième axe 0,2 %. Ces valeurs indiquent que l'essentiel de l'information est lié au premier axe. Il n'est donc pas surprenant que l'interprétation des autres axes soulève des difficultés, comme nous l'avons signalé au paragraphe 2.5.

Comme pour l'analyse en composantes principales, des critères empiriques peuvent aussi orienter le choix du nombre d'axes à retenir. Parmi ces critères, on peut citer la recherche d'un coude dans le diagramme des valeurs propres successives et l'élimination des axes relatifs aux valeurs propres situées après ce coude, ou encore l'élimination des axes relatifs aux valeurs propres inférieures à la moyenne. Ces critères sont décrits, pour l'analyse en composantes principales, notamment par Palm [1998].

Des éléments tels que la possibilité d'interprétation des axes et la qualité de la représentation des points dans le sous-espace qui est retenu entrent également en ligne de compte. Dans ce contexte, les informations fournies par les éventuelles variables supplémentaires peuvent aussi donner des informations quant à l'intérêt des facteurs. Si les modalités d'une variable supplémentaire ont des valeurs-tests très significatives sur un axe donné, l'axe présentera sans doute un intérêt du point de vue de l'interprétation.

Des techniques de rééchantillonnage (*bootstrap* et *jackknife*) peuvent également être utilisées pour vérifier la stabilité des résultats sur les différents axes, afin d'identifier et de ne retenir que les facteurs stables [Lebart *et al.*, 1995].

Enfin, dans le cas de l'analyse des correspondances simples réalisée à partir du tableau de contingence, et pour autant que les données proviennent d'un échantillon aléatoire et simple prélevé dans une très grande population, une procédure basée sur l'utilisation des variables χ^2 permet de tester, de manière approximative, la signification des facteurs [Saporta, 1990; Lebart *et al.*, 1995].

4.3. Codage des variables

L'analyse des correspondances ne traite que des variables de type qualitatif. En présence de variables quantitatives, deux solutions sont possibles. Ou bien

la variable quantitative est transformée en une variable qualitative à modalités ordonnées, ou bien la variable est traitée sous forme de variable supplémentaire, comme nous l'avons signalé au paragraphe 4.1.

La transformation de la variable quantitative en variable qualitative donne lieu à une perte de précision des observations. Elle permet, par contre, de traduire des relations non linéaires entre les variables quantitatives. Cette propriété conduit par exemple VOLLE [1985] à préconiser l'utilisation de l'analyse des correspondances multiples, pour compléter les résultats d'une analyse en composantes principales, lorsqu'on ne dispose que de variables quantitatives.

Dans le cas d'un découpage en classes d'une variable quantitative pour la rendre qualitative, le nombre de classes et les limites de classes devront être définies en se rappelant les principes présentés au paragraphe 2.2 :

- l'inertie d'un caractère est une fonction croissance du nombre de modalités;
- l'inertie d'une modalité est d'autant plus grande que la modalité est peu représentée.

Il en résulte qu'on constituera, de préférence, des modalités d'effectifs semblables et qu'on découpera les variables de manière à obtenir un nombre de modalités comparables aux nombres de modalités des autres caractères.

Ces principes ne doivent cependant pas être appliqués de manière aveugle. C'est ainsi qu'il y a lieu de respecter des seuils naturels qui existent éventuellement dans le contexte de l'étude. A titre d'exemple, dans une enquête sur les dépenses des ménages où une variable quantitative concernerait les dépenses pour la voiture, la valeur zéro attribuée par les ménages sans voiture ne devrait pas être regroupée avec des dépenses très faibles, sous prétendre d'équilibrer au mieux les effectifs des modalités.

Enfin, signalons encore que le problème du codage concerne aussi le cas de variables qualitatives. En effet, il est parfois utile de fusionner des modalités pour réduire le nombre de modalités de certains caractères ou pour équilibrer les effectifs des modalités.

Pour l'exemple traité, l'application des principes ci-dessus nous a conduit à modifier quelque peu le codage des variables initiales avant de les soumettre à l'analyse des correspondances (paragraphe 2 et annexe).

5. CONCLUSIONS

L'analyse factorielle des correspondances multiples est une extension de l'analyse des correspondances simples, par la prise en compte de plus de deux caractères. La méthode est donc spécialement destinée au traitement de tables de contingence à plus de deux dimensions.

L'extension de l'analyse des correspondances simples à l'analyse des correspondances multiples repose sur l'équivalence entre l'analyse des correspondances

du tableau de contingence et l'analyse des correspondances du tableau disjonctif complet, cette dernière analyse pouvant se généraliser au cas de plus de deux variables nominales. Cette équivalence ne doit cependant pas faire perdre de vue la spécificité de l'interprétation de certains résultats liés à l'analyse du tableau disjonctif complet.

L'analyse des correspondances multiples est une méthode dont le champ d'application est relativement large en ce qui concerne la nature des données puisqu'elle peut s'appliquer tant à des données qualitatives que quantitatives ou à un mélange de variables de nature différente, les variables quantitatives étant découpées préalablement en classes. A ce sujet, on se rappellera que ce découpage n'appauvrit pas nécessairement l'information disponible car il permet de prendre en compte d'éventuelles liaisons non linéaires.

Par ailleurs, la séparation entre variables actives et variables passives, bien qu'elle ne soit pas propre à l'analyse des correspondances, élargit encore les potentialités de la méthode. Cette dichotomie est particulièrement intéressante dans les dépouillement d'enquêtes sociologiques où il est d'usage de considérer comme variables actives celles qui décrivent plus ou moins objectivement un individu (profession, âge, sexe, catégorie socio-professionnelle, etc.) et comme variables passives les questions constituant le sujet même de l'enquête [Bouroche et Saporta, 1980].

Enfin, même lorsqu'elle ne révèle que des relations entre caractères qui pouvaient être attendues, au moins par les spécialistes du domaine, l'analyse des correspondances multiples a l'avantage de condenser l'information et de présenter une partie importante de celle-ci sous la forme de quelques graphiques assez explicites. Ce point est surtout intéressant lorsqu'on se trouve en présence d'un grand nombre de modalités et d'individus. L'analyse des correspondances multiples se révélant dans ce cas être un outil descriptif particulièrement efficace. Mais, comme le signale Dervin [1988], "les hypothèses déduites (différences entre des groupes, oppositions entre modalités de différentes variables, par exemple) devront, le plus souvent être vérifiées ultérieurement sur d'autres jeux de données par des analyses plus fines, employant des tests statistiques".

BIBLIOGRAPHIE

Benzécri J.P. [1979]. Sur le calcul des taux d'inertie dans l'analyse d'un questionnaire. *Cahiers de l'analyse des données* 4, 377-378.

Bouroche J.M., Saporta G. [1980]. *L'analyse des données.* Paris, Presses universitaires de France, 127 p.

Dervin C. [1988]. *Comment interpréter les résultants d'une analyse factorielle des correspondances?* Paris, Institut technique des Céréales et des Fourrages, 75 p.

Escofier B., Pagès J. [1998]. *Analyses factorielles simples et multiples : objectifs, méthodes et interprétation.* Paris, Dunod, 284 p.

Hosmer D.W., Lemeshow S. [1989]. *Applied logistic regression.* New York, Wiley, 307 p.

LEBART L., MORINEAU A., PIRON M. [1995]. *Statistique exploratoire multidimensionnelle*. Paris, Dunod, 439 p.

PALM R. [1993]. Les méthodes d'analyse factorielle : principes et applications. *Notes Stat. Inform.* (Gembloux), 93/1, 38 p. (réédition en 2002).

PALM R. [1998]. L'analyse en composantes principales : principes et applications. *Notes Stat. Inform.* (Gembloux), 98/2, 31 p.

SAPORTA G. [1990]. *Probabilités, analyse des données et statistique*. Paris, Technip, 493 p.

SAS INSTITUTE INC [1989]. *User's guide, version 6*, Fourth edition (2 volumes). Cary NC, SAS Institute Inc., 943 + 846 p.

TOMASSONE R., DERVIN C., MASSON J.P. [1993]. *Biométrie : modélisation de phénomènes biologiques*. Paris, Masson, 553 p.

VOLLE M. [1985]. *Analyse des données*. Paris, Economica, 324 p.

ANNEXE

```
DATA MAMMOG;
/*
 Origine : HOSMER W. et LEMESHOW S. (1989)
           Appendix 6 p. 279 et ss
*/
INPUT OBS x1-x6;
if X1=1 then  MAMMO='M1AN';
  else if X1=2 then  MAMMO='MP1AN';
  else  MAMMO='MJAM';
if X2=1 or X2=2 then  SYMPT='SYMPT1';
  else if X2=3 then SYMPT='SYMPT2';
  else SYMPT='SYMPT3';
if X3=5 or X3=6 then  INTERET='INTGRAN';
  else if X3=7 or X3=8 then  INTERET='INTMOY';
  else  INTERET='INTFAIB';
if X4=0 then  ANTEC='ANTNON';
  else ANTEC='ANTOUI';
if X5=0 then EXASEIN='EXANON';
  else  EXASEIN='EXAOUI';
if X6=1  then  PDETEC='PDETGRAN';
  else if X6=2  then  PDETEC='PDETMOY'  ;
  else   PDETEC='PDETFAIB'  ;
drop x1-x6;
CARDS;
  1  3  3   7  0  1  2
  2  3  2  11  0  1  1
  3  3  3   8  1  1  1
  4  1  3  11  0  1  1
  5  2  4   7  0  1  1

 (suite des données)

407  3  3  11  0  1  2
408  1  3  10  0  1  2
409  3  4   8  0  1  1
410  1  4   6  0  1  1
411  2  2   6  0  1  1
412  3  4   7  0  1  1
;

PROC CORRESP DATA=MAMMOG OUT=SCORES MCA OBSERVED ;
  TABLES MAMMO SYMPT INTERET ANTEC EXASEIN PDETEC ;

PROC PLOT DATA=SCORES;
  PLOT DIM2*DIM1='*' $ _name_ /HREF=0 VREF=0;

run;
```