NOTES DE STATISTIQUE
ET D'INFORMATIQUE

93/1

LES METHODES D'ANALYSE FACTORIELLE: PRINCIPES ET APPLICATIONS

R. PALM

Faculté universitaire des Sciences agronomiques
GEMBLOUX
(Belgique)

LES MÉTHODES D'ANALYSE FACTORIELLE : PRINCIPES ET APPLICATIONS

R. Palm *

1. RÉSUMÉ

Cette note présente de façon simple les principes de base des méthodes d'analyse factorielle, et plus particulièrement de l'analyse en composantes principales et de l'analyse factorielle des correspondances. L'accent est mis sur les éléments utiles lors de l'interprétation des résultats et les méthodes sont illustrées par des exemples élémentaires, traités de faèon détaillée.

2. SUMMARY

This note describes in a simple way the principles of the factorial methods, and more particularly of the principal component analysis and the correspondence analysis. The emphasis is put on the elements useful when interpreting the results of an analysis and the methods are illustrated by completely worked out examples.

3. INTRODUCTION

Les méthodes d'analyse factorielle sont incontestablement des outils fondamentaux de l'analyse des tableaux de données qui ne présentent pas de structure particulière. Elles visent essentiellement un but descriptif, en condensant l'information contenue dans un tableau, constitué souvent d'un nombre élevé de lignes et de colonnes, en quelques représentations graphiques à deux dimensions, accompagnées de tableaux reprenant les valeurs numériques de caractéristiques destinées à aider l'utilisateur lors de l'interprétation.

L'analyse en composantes principales et l'analyse factorielle des correspondances sont sans doute les deux méthodes factorielles les plus couramment utilisées, mais d'autres méthodes peuvent encore être envisagées.

*Chef de travaux et Maître de conférences à la Faculté des Sciences agronomiques de Gembloux (Unité de Statistique et Informatique).

L'objectif de cette note est de décrire de façon simple et d'illustrer par des exemples élémentaires, les principes de base de ces deux méthodes. La bonne compréhension de ces principes devrait permettre aux utilisateurs d'interpréter correctement les documents imprimés fournis par les logiciels statistiques. Une attention plus particulière est accordée à l'analyse des correspondances, car l'analyse en composantes principales présente moins de difficultés et est en outre exposée de faèon claire et simple par DAGNELIE [1982]. En particulier, pour cette dernière méthode nous ne présentons pas de documents imprimés fournis par les logiciels. Pour l'analyse des correspondances, par contre, nous reproduisons plusieurs figures obtenues par la procédure CORRESP du logiciel SAS. Des informations relatives à cette procédure sont données dans le manuel d'utilisation de ce logiciel [X, 1990].

L'analyse en composantes principales et l'analyse des correspondances présentent un certain nombre de points communs et peuvent être considérées comme des applications particulières d'une méthode générale dont nous exposons les principes au paragraphe 2. Nous examinons ensuite l'analyse en composantes principales (paragraphe 3) et l'analyse des correspondances (paragraphe 4), puis nous terminons par quelques informations complémentaires (paragraphe 5).

Dans cette note, nous nous limitons à une présentation relativement élémentaire des méthodes. En particulier, aucune démonstration mathématique n'est donnée. Le lecteur souhaitant approfondir ces méthodes trouvera des informations plus détaillées dans l'abondante littérature consacrée à ce sujet, par exemple dans BENZÉCRI et BENZÉCRI [1980], BOUROCHE et SAPORTA [1980], CIBOIS [1983], FÉNELON [1981], GREENACRE [1984], JACKSON [1991] et LEBART *et al.* [1979].

4. PRINCIPES GÉNÉRAUX

4.1. Valeurs propres, vecteurs propres et reconstitution du tableau de données

Soit une matrice X de dimensions $n \times p$, telles que $p \leqslant n$, contenant des nombres quelconques. Considérons d'abord la matrice carrée $\boldsymbol{X}'\boldsymbol{X}$, de dimensions $p \times p$ et de rang r ($r \leqslant p$). Cette matrice admet r valeurs propres positives :

$$\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_r \,,$$

auxquelles sont associés r vecteurs propres $\boldsymbol{u_1}$, $\boldsymbol{u_2}$, $\ldots$, $\boldsymbol{u_r}$.

Considérons maintenant la matrice carrée $\boldsymbol{X}\boldsymbol{X}'$, de dimensions $n \times n$. Cette matrice est également de rang r et possède r valeurs propres positives :

$$\mu_1 \geqslant \mu_2 \geqslant \ldots \geqslant \mu_r \,,$$

auxquelles sont associés r vecteurs propres $\boldsymbol{v_1}$, $\boldsymbol{v_2}$, $\ldots$, $\boldsymbol{v_r}$.

On peut montrer que les valeurs propres non nulles sont identiques pour les deux matrices :

$$\lambda_k = \mu_k \qquad (k = 1, \ldots, r) \,,$$

et que les vecteurs propres des deux matrices sont liés par les relations suivantes :

$$\boldsymbol{u_k} = \frac{1}{\sqrt{\lambda_k}}\, \boldsymbol{X}'\, \boldsymbol{v_k} \qquad (k = 1, \ldots, r)$$

et

$$\boldsymbol{v_k} = \frac{1}{\sqrt{\lambda_k}}\, \boldsymbol{X}\, \boldsymbol{u_k} \qquad (k = 1, \ldots, r)\,.$$

A titre d'illustration, considérons la matrice suivante, de dimensions 4×3 :

$$\boldsymbol{X} = \begin{bmatrix} 15 & 5 & 11 \\ 8 & 1 & 6 \\ 10 & 6 & 8 \\ 19 & 14 & 14 \end{bmatrix}.$$

On en déduit :

$$\boldsymbol{X}'\,\boldsymbol{X} = \begin{bmatrix} 750 & 409 & 559 \\ 409 & 258 & 305 \\ 559 & 305 & 417 \end{bmatrix} \quad \text{et} \quad \boldsymbol{X}\,\boldsymbol{X}' = \begin{bmatrix} 371 & 191 & 268 & 509 \\ 191 & 101 & 134 & 250 \\ 268 & 134 & 200 & 386 \\ 509 & 250 & 386 & 753 \end{bmatrix}.$$

Les valeurs propres non nulles sont :

$$\lambda_1 = \mu_1 = 1.395,59\,, \quad \lambda_2 = \mu_2 = 29,18 \quad \text{et} \quad \lambda_3 = \mu_3 = 0,23\,.$$

Les vecteurs propres associés sont, pour la matrice $\boldsymbol{X}'\boldsymbol{X}$:

$$\boldsymbol{u_1} = \begin{bmatrix} 0,7315 \\ 0,4092 \\ 0,5454 \end{bmatrix}, \quad \boldsymbol{u_2} = \begin{bmatrix} -0,3291 \\ 0,9124 \\ -0,2433 \end{bmatrix}, \quad \boldsymbol{u_3} = \begin{bmatrix} 0,5972 \\ 0,0015 \\ -0,8021 \end{bmatrix},$$

et pour la matrice $\boldsymbol{X}\,\boldsymbol{X}'$:

$$\boldsymbol{v_1} = \begin{bmatrix} 0,5091 \\ 0,2552 \\ 0,3783 \\ 0,7298 \end{bmatrix}, \quad \boldsymbol{v_2} = \begin{bmatrix} -0,5646 \\ -0,5887 \\ 0,0440 \\ 0,5768 \end{bmatrix}, \quad \boldsymbol{v_3} = \begin{bmatrix} 0,2965 \\ -0,0698 \\ -0,9078 \\ 0,2882 \end{bmatrix}.$$

Ces résultats ont été obtenus à l'aide du logiciel Minitab, en utilisant les commandes relatives aux opérations sur les matrices. Pour plus de clarté, nous avons conservé uniquement quatre décimales, mais pour la suite des calculs, nous prendrons cependant en considération les valeurs en mémoire dans l'ordinateur, afin d'éviter de perdre trop de précision. Par ailleurs, on sait également que les vecteurs propres ont un signe arbitraire lié à l'algorithme de calcul utilisé et

qu'on peut changer le signe de tous les éléments d'un vecteur propre. C'est ce qui a été fait pour les vecteurs $\boldsymbol{u_1}$ et $\boldsymbol{u_2}$, afin de garantir l'égalité :

$$\boldsymbol{u_k} = \frac{1}{\sqrt{\lambda_k}}\, \boldsymbol{X}'\boldsymbol{v_k} \qquad (k = 1,\ 2 \text{ ou } 3)\,.$$

A partir des valeurs propres et des vecteurs propres des matrices $\boldsymbol{X}'\,\boldsymbol{X}$ et $\boldsymbol{X}\,\boldsymbol{X}'$, on peut reconstituer la matrice $\boldsymbol{X}$, par la relation suivante :

$$\boldsymbol{X} = \sum_{k=1}^{r} \sqrt{\lambda_k}\; \boldsymbol{v_k}\, \boldsymbol{u_k'} = \sum_{k=1}^{r} \boldsymbol{X_k}\,.$$

Chacune des matrices $\boldsymbol{X_k}$ $(k = 1, \ldots, r)$ est égale, à une constante près, au produit scalaire de deux vecteurs, et est de rang 1. La matrice $\boldsymbol{X}$, de rang r est, par conséquent, égale à la somme de r matrices de rang 1.

Pour l'exemple considéré, on a, aux erreurs d'arrondis près :

$$\boldsymbol{X} = \begin{bmatrix} 13,91 & 7,78 & 10,37 \\ 6,97 & 3,90 & 5,20 \\ 10,34 & 5,78 & 7,71 \\ 19,94 & 11,16 & 14,87 \end{bmatrix} + \begin{bmatrix} 1,00 & -2,78 & 0,74 \\ 1,05 & -2,90 & 0,77 \\ -0,08 & 0,22 & -0,06 \\ -1,03 & 2,84 & -0,76 \end{bmatrix}$$
$$+ \begin{bmatrix} 0,09 & 0,00 & -0,11 \\ -0,02 & 0,00 & 0,03 \\ -0,26 & 0,00 & 0,35 \\ 0,08 & 0,00 & -0,11 \end{bmatrix}.$$

On constate que les trois matrices, $\boldsymbol{X_1}$, $\boldsymbol{X_2}$ et $\boldsymbol{X_3}$, qui permettent de reconstituer $\boldsymbol{X}$, sont d'importance décroissante. On constate aussi que la première matrice constitue déjà une très bonne approximation de $\boldsymbol{X}$ et qu'une approximation encore meilleure est donnée par la somme des deux premières matrices. Pour préciser la qualité de ces deux approximations, on peut calculer la somme des carrés des écarts entre les valeurs de départ et les valeurs reconstituées à partir de la première matrice ou de la somme des deux premières matrices. On sait que la somme des carrés des éléments de la matrice $\boldsymbol{X}$, qui est aussi la trace de $\boldsymbol{X}'\,\boldsymbol{X}$ et de $\boldsymbol{X}\,\boldsymbol{X}'$, est égale à la somme des valeurs propres, soit 1.425. Par ailleurs, on peut vérifier que les sommes des carrés des éléments des matrices $\boldsymbol{X_1}$, $\boldsymbol{X_2}$ et $\boldsymbol{X_3}$ sont égales aux valeurs propres λ_1, λ_2, et λ_3. De plus, la somme des carrés des éléments de la somme de deux ou plusieurs matrices servant à reconstituer $\boldsymbol{X}$ est égale à la somme des valeurs propres correspondantes. Par conséquent, la somme des carrés des écarts entre $\boldsymbol{X}$ et $\boldsymbol{X_1}$ est égale à la somme des carrés des éléments de $\boldsymbol{X_2} + \boldsymbol{X_3}$, soit :

$$\lambda_2 + \lambda_3 = 29,18 + 0,23 = 29,41\,,$$

et la somme des carrés des écarts entre $\boldsymbol{X}$ et $(\boldsymbol{X_1} + \boldsymbol{X_2})$ est égale à la somme des carrés des éléments de $\boldsymbol{X_3}$, c'est-à-dire aussi à λ_3, soit 0,23. Si on exprime ces résultats en pour cent de la somme des carrés des éléments de $\boldsymbol{X}$, on peut dire que $\boldsymbol{X_1}$ est une approximation qui reconstitue 97,94 % de $\boldsymbol{X}$ et que l'erreur

d'approximation est de 2,06 %, et on peut dire aussi que $(\boldsymbol{X_1} + \boldsymbol{X_2})$ reconstitue 99,98 % de $\boldsymbol{X}$ et que l'erreur d'approximation est dans ce cas de 0,02 %.

La matrice $\boldsymbol{X_1}$ est en fait la matrice résultant du produit scalaire de deux vecteurs $\boldsymbol{v_1}$ et $\boldsymbol{u_1'}$ qui minimise la somme des carrés des écarts entre les valeurs de $\boldsymbol{X}$ et de $\boldsymbol{X_1}$. De même, $\boldsymbol{X_2}$ est la matrice résultant du produit scalaire de deux vecteurs, $\boldsymbol{v_2}$ et $\boldsymbol{u_2'}$, $\boldsymbol{v_2}$ étant orthogonal à $\boldsymbol{v_1}$ et $\boldsymbol{u_2}$ étant orthogonal à $\boldsymbol{u_1}$, qui minimise la somme des carrés des écarts entre les valeurs de $(\boldsymbol{X} - \boldsymbol{X_1})$ et de $\boldsymbol{X_2}$. Enfin, $\boldsymbol{X_3}$ est aussi une matrice résultant du produit scalaire de deux vecteurs, $\boldsymbol{v_3}$ et $\boldsymbol{u_3'}$, qui reconstitue exactement la matrice des écarts $\boldsymbol{X} - (\boldsymbol{X_1} + \boldsymbol{X_2})$.

Ce que nous avons vérifié sur l'exemple numérique peut évidemment se généraliser à toute matrice $\boldsymbol{X}$ de dimensions $n \times p$. Si la matrice $\boldsymbol{X'X}$ possède r valeurs propres différentes de zéro, la matrice $\boldsymbol{X}$ est égale à la somme de r matrices, chacune d'elles étant obtenue à partir du produit d'un vecteur propre de $\boldsymbol{X'X}$ et du vecteur propre correspondant de $\boldsymbol{XX'}$. L'importance de ces différentes matrices dans la reconstitution de $\boldsymbol{X}$ est proportionnelle à la valeur propre relative aux vecteurs propres qui sont multipliés.

L'intérêt de cette reconstitution réside surtout dans l'approximation de $\boldsymbol{X}$ au moyen de q matrices de rang 1, q étant inférieur au nombre de valeurs propres non nulles. Une bonne approximation d'un tableau de $n \times p$ nombres peut, en effet, souvent être obtenue à partir d'un nombre réduit de vecteurs propres.

4.2. Interprétation géométrique et représentation graphique

A la décomposition de la matrice $\boldsymbol{X}$ en une somme de matrices de rang 1 correspond également une interprétation géométrique.

On sait que la matrice $\boldsymbol{X}$ peut être représentée par n vecteurs-lignes dans l'espace des p colonnes ou par p vecteurs-colonnes dans l'espace des n lignes [Dagnelie, 1982]. Considérons d'abord l'espace à p dimensions. Une ligne quelconque i $(i = 1, \ldots, n)$ de $\boldsymbol{X}$ est représentée dans cet espace par un point dont les p coordonnées sur les p axes de l'espace sont les éléments de la ligne i.

Les éléments du premier vecteur propre de $\boldsymbol{X'X}$ sont les coefficients directeurs de l'axe qui maximise la somme des carrés des projections des n points sur l'axe en question. Cet axe est appelé èpremier axe factorielè. Les projections de ces points sur l'axe sont égales à $\boldsymbol{Xu_1}$ et la somme des carrés de ces projections est égale à la première valeur propre, λ_1, de $\boldsymbol{X'X}$. De même, les éléments du deuxième vecteur propre de $\boldsymbol{X'X}$ sont les coefficients directeurs de l'axe 2, perpendiculaire à l'axe 1 défini précédemment, qui maximise la somme des carrés des projections des n points sur l'axe 2. Les projections sur ce deuxième axe factoriel sont égales à $\boldsymbol{X}\,\boldsymbol{u_2}$ et la somme des carrés de ces projections est égale à la deuxième valeur propre, λ_2, de $\boldsymbol{X'X}$.

Il faut noter aussi que l'espace à deux dimensions formé par les axes 1 et 2 est le plan qui maximise la somme des carrés des distances à l'origine des axes des projections des n points sur le plan. Mais c'est aussi l'espace qui miniminise la somme des carrés des distances entre les n points dans l'espace à p dimensions et leur projection sur le plan formé par l'axe 1 et l'axe 2.

Des interprétations analogues peuvent être données aux autres vecteurs propres, $\boldsymbol{u_3}$, ..., $\boldsymbol{u_p}$, les axes engendrés par ces vecteurs étant chaque fois perpendiculaires à l'ensemble des axes déjà définis. Toutefois, si le rang r de la matrice $\boldsymbol{X}$ est plus petit que p, les axes $r+1$, $r+2$, ..., p, ne présentent aucun intérêt, puisque la projection des n points sur ces axes vaut zéro, la somme des carrés des projections étant, en effet, égale à la valeur propre correspondante.

On peut évidemment s'intéresser aussi à la représentation géométrique des p colonnes dans l'espace des n lignes et construire n axes perpendiculaires deux à deux dont les coefficients directeurs sont les composantes des n vecteurs propres de la matrice $\boldsymbol{X}\,\boldsymbol{X}'$. Les projections des p points sur ces axes sont égales à $\boldsymbol{X}'\,\boldsymbol{v_k}$, et la somme des carrés des projections est égale à λ_k. Les $n-r$ derniers axes sont cependant sans intérêt, la projection des points sur ces axes étant nulle.

Nous avons vu précédemment qu'une reconstitution approchée de la matrice $\boldsymbol{X}$ pouvait être obtenue en ne prenant en considération que les q premiers vecteurs propres ($q \leqslant r$). D'un point de vue géométrique, cela revient à remplacer l'espace à p (ou n) dimensions par un sous-espace à q dimensions et à considérer que les n (ou p) points se situent exactement dans cet espace. La qualité de cette approximation est donnée par le rapport :

$$\sum_{k=1}^{q} \lambda_k \Big/ \sum_{k=1}^{r} \lambda_k \,,$$

qui mesure le rapport entre la somme des carrés des distances des points projetés dans l'espace à q dimensions et la somme des carrés des distances des points dans l'espace à p (ou n) dimensions.

Ce rapport ne donne cependant qu'une idée globale de la qualité de la représentation pour l'ensemble des points, et il peut y avoir des différences importantes d'un point à l'autre, un point pouvant, par exemple, être très près du sous-espace à q dimensions et l'autre au contraire pouvant être plus éloigné de ce sous-espace. Pour chiffrer la qualité de la représentation d'un point donné, on peut calculer, par exemple, le rapport entre le carré de la distance à l'origine de la projection de ce point dans le sous-espace à q dimensions et le carré de la distance à l'origine de ce point dans l'espace à p ou à n dimensions.

La qualité de l'approximation d'un point par un axe donné mesure en fait l'importance de l'angle formé par le segment reliant l'origine des axes au point donné et la projection de ce segment sur l'axe en question. De faèon plus précise, il s'agit du carré du cosinus de cet angle. Une interprétation géométrique analogue peut être donnée à la qualité de l'approximation d'un point par un ensemble d'axes : il s'agit alors du carré du cosinus de l'angle formé par le segment reliant l'origine des axes au point donné et la projection de ce segment dans l'espace des axes pris en considération. Cette valeur est d'ailleurs égale à la somme des carrés des cosinus de la projection sur chacun des axes.

Une représentation schématique de la qualité des approximations est donnée, dans le cas de deux axes, à la figure 1 : la qualité de l'approximation du point P par l'axe 1 est égale à $\cos^2 \alpha_1$; la qualité de l'approximation de ce point

par l'axe 2 est égale à $\cos^2 \alpha_2$ et la qualité de l'approximation du point P par la plan formé par les axes 1 et 2 est égale à :

$$\cos^2 \alpha = \cos^2 \alpha_1 + \cos^2 \alpha_2 \,.$$

Figure 1. Qualité de l'approximation d'un point par un axe ou un plan (P_1 = projection du point P sur l'axe 1 ; P_2 = projection du point P sur l'axe 2 ; P_{12} = projection du point P sur le plan formé par les axes 1 et 2).

On notera que, dans la littérature statistique, la qualité de l'approximation d'un point est souvent appelée ècosinus carréè, du fait de l'interprétation géométrique qui vient d'en être donnée, ou encore ècontribution d'un axe à l'explication d'un pointè ou ècontribution relative d'un axe à la position d'un pointè.

A titre d'illustration, reprenons la matrice $\boldsymbol{X}$ définie au paragraphe précédent et calculons les coordonnées des quatre points-lignes et des trois points-colonnes sur les axes factoriels. On trouve :

$$\boldsymbol{X}\,\boldsymbol{u_1} = \begin{bmatrix} 19,02 \\ 9,53 \\ 14,13 \\ 27,26 \end{bmatrix}, \quad \boldsymbol{X}\,\boldsymbol{u_2} = \begin{bmatrix} -3,05 \\ -3,18 \\ 0,24 \\ 3,12 \end{bmatrix} \quad \text{et} \quad \boldsymbol{X}\,\boldsymbol{u_3} = \begin{bmatrix} 0,14 \\ -0,03 \\ -0,44 \\ 0,14 \end{bmatrix},$$

$$\boldsymbol{X'}\,\boldsymbol{v_1} = \begin{bmatrix} 27,33 \\ 15,29 \\ 20,37 \end{bmatrix}, \quad \boldsymbol{X'}\,\boldsymbol{v_2} = \begin{bmatrix} -1,78 \\ 4,93 \\ -1,31 \end{bmatrix} \quad \text{et} \quad \boldsymbol{X'}\,\boldsymbol{v_3} = \begin{bmatrix} 0,29 \\ 0,00 \\ -0,39 \end{bmatrix}.$$

On vérifie bien que la somme des carrés des projections des points-lignes ou des points-colonnes sur chacun des axes est égale, aux erreurs d'arrondis près, à la valeur propre associée à l'axe. Ainsi, pour le premier axe par exemple, on a :

$$19,02^2 + 9,53^2 + 14,13^2 + 27,26^2 = 1.395,35$$

et

$$27,33^2 + 15,29^2 + 20,37^2 = 1.395,65\,.$$

La figure 2 donne une représentation graphique des matrices $\boldsymbol{X}$, $\boldsymbol{X_1}$ et $\boldsymbol{X_2}$ dans l'espace des colonnes. Chaque matrice est représentée par quatre points dans l'espace à trois dimensions, appelées x, y et z. Les points relatifs à $\boldsymbol{X}$ sont représentés par des cercles ; les points relatifs à $\boldsymbol{X_1}$ par des carrés et les points relatifs à $\boldsymbol{X_2}$ par des losanges.

On vérifie bien que les quatre carrés sont sur une droite, qui correspond au premier axe factoriel. Les écarts entre les cercles et les carrés sont assez faibles, car $\boldsymbol{X_1}$ est une bonne approximation de $\boldsymbol{X}$, puisqu'elle reconstitue 97,94 % de $\boldsymbol{X}$ (paragraphe 2.1). Mais on voit également que la qualité de la reconstitution n'est pas identique pour les quatre points. L'écart est du même ordre de grandeur pour les points correspondant aux lignes 1, 2 et 4, mais il est beaucoup plus faible pour le point correspondant à la ligne 3. Les valeurs des cosinus carrés confirment la

Figure 2. Représentation des points-lignes dans l'espace des colonnes (x, y et z) et projections des points-lignes sur les deux premiers axes factoriels (les cercles représentent les points-lignes dans l'espace des colonnes, les carrés correspondent aux projections sur le premier axe factoriel et les losanges correspondent aux projections sur le deuxième axe factoriel).

bonne approximation de la position du point relatif à la troisième ligne de $\boldsymbol{X}$ par sa projection sur le premier axe. On a en effet :

$$\cos^2 = 14,13^2/(14,13^2 + 0,24^2 + 0,44^2) = 0,9987 \text{ ou } 99,87\,\%\,,$$

pour la ligne 3, alors que pour les trois autres lignes, on obtient les cosinus carrés suivants : 97,49 %, 89,98 % et 98,70 %. On note aussi que le cosinus carré relatif à la ligne 2 est nettement plus faible que les cosinus carrés relatifs aux lignes 1 et 4, alors que l'écart entre les cercles et les carrés est du même ordre de grandeur. Cela s'explique par les distances par rapport à l'origine de ces points, qui est plus faible pour la ligne 2 que pour les autres lignes. D'autre part, l'examen des écarts entre les cercles et les carrés montre que ces écarts se manifestent surtout selon la coordonnée y. Il n'est donc pas étonnant que, dans $\boldsymbol{X_2}$, ce soit précisément la seconde colonne qui présente les valeurs les plus différentes de zéro.

Tout comme pour $\boldsymbol{X_1}$, on voit que les points relatifs à $\boldsymbol{X_2}$, représentés par des losanges, se situent également sur une droite, qui correspond au deuxième axe factoriel et qui est perpendiculaire au premier axe. Par eux-mêmes, ces points ne présentent guère d'intérêt, mais ils permettent de définir les coordonnées, dans le système (x, y, z), des projections des points relatifs à $\boldsymbol{X}$ dans le sous-espace des deux premiers axes factoriels. Ces projections sont les points relatifs à la matrice $(\boldsymbol{X_1} + \boldsymbol{X_2})$. Nous n'avons pas représenté ces projections sur la figure, car elles sont pratiquement confondues avec les points relatifs à $\boldsymbol{X}$, la qualité globale de l'approximation par les deux axes étant égale à :

$$(1.395,59 + 29,18)/1.425 = 0,9998 \text{ ou } 99,98\,\%\,.$$

La qualité de l'approximation pour le premier point est égale à :

$$(19,02^2 + 3,05^2)/(19,02^2 + 3,05^2 + 0,14^2) = 0,9999 \text{ ou } 99,99\,\%\,,$$

et pour les trois autres points, on trouve respectivement, 100,00 %, 99,90 % et 100,00 %, aux erreurs d'arrondis près. La qualité de la représentation des points-lignes dans un espace à deux dimensions est donc excellente pour tous les points.

La représentation graphique de la matrice $\boldsymbol{X}$ dans l'espace des quatre lignes ne peut évidemment pas être réalisée, puisqu'il s'agit d'un espace à quatre dimensions. Par contre, on pourrait représenter les trois points-colonnes dans le sous-espace constitué par les trois axes factoriels. Plus simplement encore, on pourrait projeter les trois points sur le plan formé par les deux premiers axes factoriels, ce qui donnerait une bonne image de la position relative des trois

points-colonnes correspondant à $\boldsymbol{X}$, puisque la qualité de la représentation des trois points dans ce sous-espace vaut, respectivement, 99,99 %, 100,00 % et 99,96 %, ce qui, une fois encore, confirme le rôle négligeable de $\boldsymbol{X_3}$ dans la reconstitution de $\boldsymbol{X}$ par la relation :

$$\boldsymbol{X} = \boldsymbol{X_1} + \boldsymbol{X_2} + \boldsymbol{X_3}\,,$$

décrite au paragraphe 2.1.

En pratique, les représentations graphiques analogues à la figure 2 ne sont jamais établies, car les nombres de lignes et de colonnes de $\boldsymbol{X}$ sont le plus souvent supérieurs à deux. Par contre les représentations des projections sur les différents plans factoriels sont à la base des interprétations des résultats des analyses factorielles, comme nous l'avons déjà signalé dans l'introduction (paragraphe 1) et comme nous le verrons encore dans les paragraphes suivants.

5. ANALYSE EN COMPOSANTES PRINCIPALES

5.1. Transformation des données

Considérons maintenant qu'on dispose des observations relatives à p variables et effectuées sur n individus. Ces données peuvent se représenter sous la forme d'une matrice, $\boldsymbol{Y}$, de dimensions $n \times p$. Les p variables sont le plus souvent de nature différente et une analyse de la matrice telle que nous venons de la décrire au paragraphe précédent ne présenterait guère d'intérêt. En effet, un simple changement d'origine ou d'unités d'une ou de plusieurs variables conduirait à une modification des résultats. Pour cette raison, la matrice de données $\boldsymbol{Y}$ sera généralement transformée d'abord en une matrice $\boldsymbol{X}$ par la relation suivante :

$$x_{ij} = (y_{ij} - \bar{y}_j)/(s_j\sqrt{n}) \qquad (i = 1, \ldots, n\,;\; j = 1, \ldots, p)\,.$$

Dans cette relation, $\bar{y}_j$ et s_j sont, respectivement, la moyenne arithmétique et l'écart-type observé de la colonne j :

$$\bar{y}_j = \sum_{i=1}^{n} y_{ij} \quad \text{et} \quad s_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{ij} - \bar{y}_j)^2} \qquad (j = 1, \ldots, p)\,.$$

On remplace donc les variables initiales par les variables centrées réduites correspondantes, multipliées par le facteur $1/\sqrt{n}$, dont le rôle est de faire coïncider la matrice $\boldsymbol{X'X}$ avec la matrice de corrélation. Dans certaines situations particulières (variables exprimées dans les mêmes unités et avec des écarts-types du même ordre de grandeur, par exemple), on ne divise pas les écarts par rapport à la moyenne par l'écart-type ; dans ces cas, la matrice $\boldsymbol{X'X}$ est la matrice des variances et covariances observées.

La matrice $\boldsymbol{X}$ peut faire l'objet de l'analyse décrite au paragraphe 2. Une telle analyse porte alors le nom d'analyse en composantes principales [2].

2. En anglais : *principal component analysis*.

A titre d'illustration, considérons la matrice :

$$\boldsymbol{Y} = \begin{bmatrix} 1.355 & 13,7 & 6,24 \\ 1.412 & 9,4 & 7,78 \\ 1.242 & 10,9 & 6,94 \\ 1.364 & 9,1 & 7,56 \end{bmatrix},$$

qui correspond aux longueurs des fibres, aux diamètres du lumen des fibres et aux épaisseurs des fibres observées sur des échantillons de bois de quatre hêtres. Les données, exprimées en microns, représentent en fait des valeurs moyennes relatives à une série de mesures effectuées sur chaque arbre et sont extraites d'un ensemble beaucoup plus vaste d'observations réalisées par Leclercq [1979]. L'analyse de ce petit tableau va nous permettre d'illustrer le mécanisme de l'analyse en composantes principales, même si une telle analyse ne présente aucun intérêt pratique dans le cas d'un tableau de dimensions aussi réduites.

A partir de la matrice $\boldsymbol{Y}$, on peut déduire la matrice $\boldsymbol{X}$, en appliquant la transformation donnée ci-dessus. On obtient :

$$\boldsymbol{X} = \begin{bmatrix} 0,0942 & 0,8030 & -0,7428 \\ 0,5514 & -0,3775 & 0,5425 \\ -0,8120 & 0,0343 & -0,1586 \\ 0,1664 & -0,4599 & 0,3589 \end{bmatrix}.$$

5.2. Valeurs propres, vecteurs propres et reconstitution du tableau de données

Les valeurs propres de la matrice de corrélation :

$$\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} 1,0000 & -0,2369 & 0,4176 \\ -0,2369 & 1,0000 & -0,9718 \\ 0,4176 & -0,9718 & 1,0000 \end{bmatrix},$$

sont égales à :

$$\lambda_1 = 2,1579\,, \quad \lambda_2 = 0,8324 \quad \text{et} \quad \lambda_3 = 0,0097\,;$$

et les vecteurs propres associés à ces valeurs propres sont :

$$\boldsymbol{u_1} = \begin{bmatrix} -0,3732 \\ 0,6400 \\ -0,6717 \end{bmatrix}, \quad \boldsymbol{u_2} = \begin{bmatrix} 0,9166 \\ 0,3663 \\ -0,1602 \end{bmatrix} \quad \text{et} \quad \boldsymbol{u_3} = \begin{bmatrix} -0,1435 \\ 0,6755 \\ 0,7233 \end{bmatrix}.$$

Quant à la matrice $\boldsymbol{X}\boldsymbol{X}'$, elle est égale à :

$$\boldsymbol{X}\,\boldsymbol{X}' = \begin{bmatrix} 1,2055 & -0,6541 & 0,0688 & -0,6202 \\ -0,6541 & 0,7408 & -0,5467 & 0,4600 \\ 0,0688 & -0,5467 & 0,6857 & -0,2078 \\ -0,6202 & 0,4600 & -0,2078 & 0,3680 \end{bmatrix}.$$

Ses trois premières valeurs propres sont égales à λ_1, λ_2 et λ_3 et la quatrième valeur propre est nulle. Les trois premiers vecteurs propres sont :

$$\boldsymbol{v_1} = \begin{bmatrix} 0,6656 \\ -0,5526 \\ 0,2937 \\ -0,4067 \end{bmatrix}, \quad \boldsymbol{v_2} = \begin{bmatrix} 0,5475 \\ 0,3071 \\ -0,7742 \\ -0,0805 \end{bmatrix} \quad \text{et} \quad \boldsymbol{v_3} = \begin{bmatrix} -0,0851 \\ 0,5919 \\ 0,2536 \\ -0,7604 \end{bmatrix}.$$

A partir des vecteurs propres $\boldsymbol{u}_k$ et $\boldsymbol{v}_k$ ($k = 1, \ldots, 3$), on peut reconstituer la matrice $\boldsymbol{X}$:

$$\boldsymbol{X} = \begin{bmatrix} -0,3648 & 0,6257 & -0,6567 \\ 0,3029 & -0,5195 & 0,5452 \\ -0,1610 & 0,2762 & -0,2898 \\ 0,2229 & -0,3824 & 0,4013 \end{bmatrix} + \begin{bmatrix} 0,4579 & 0,1830 & -0,0800 \\ 0,2568 & 0,1026 & -0,0449 \\ -0,6474 & -0,2587 & 0,1132 \\ -0,0673 & -0,0269 & 0,0118 \end{bmatrix}$$
$$+ \begin{bmatrix} 0,0012 & -0,0057 & -0,0061 \\ -0,0084 & 0,0394 & 0,0422 \\ -0,0036 & 0,0169 & 0,0181 \\ 0,0107 & -0,0506 & -0,0542 \end{bmatrix}.$$

Si on se limite à la première matrice du membre de droite, $\boldsymbol{X}_1$, on a une approximation de $\boldsymbol{X}$ telle que la somme des carrés des écarts entre les éléments de $\boldsymbol{X}$ et de $\boldsymbol{X_1}$ est minimum, sous la contrainte que la matrice $\boldsymbol{X_1}$ soit obtenue par le produit de deux vecteurs. Si on transforme la matrice $\boldsymbol{X_1}$ en une matrice $\boldsymbol{Y_1}$, en multipliant les éléments de colonne j par $s_j\sqrt{n}$ et en additionnant $\bar{y}_j$, cette matrice $\boldsymbol{Y_1}$ est celle qui minimise la somme des carrés des écarts entre valeurs observées et valeurs approchées, les écarts étant pondérés par les variances des variables s_j^2.

5.3. Représentation des individus

Pour la représentation des individus dans l'espace des variables, la transformation de $\boldsymbol{Y}$ en $\boldsymbol{X}$ revient à placer l'origine des axes au centre de gravité du nuage des points correspondant aux individus et à modifier les unités des variables. Les vecteurs propres $\boldsymbol{u_1}$, $\boldsymbol{u_2}$ et $\boldsymbol{u_3}$ sont les coefficients directeurs des axes factoriels : le premier axe factoriel minimise la somme des carrés des distances entre les points correspondant aux lignes de $\boldsymbol{X}$ et leur projection sur l'axe, le deuxième axe est perpendiculaire à l'axe 1 et minimise la somme des carrés des distances entre les points et le premier plan factoriel et, enfin, le troisième axe est perpendiculaire au premier plan factoriel. Les projections des points sur chacun de ces axes, qui sont appelées "scores" ou "valeurs des composantes principales", sont données par les produits :

$$\boldsymbol{z}_k = \boldsymbol{X}\,\boldsymbol{u_k} = \sqrt{\lambda_k}\,\boldsymbol{v_k} \qquad (k = 1, \ldots, 3)\,.$$

Elles sont égales à :

$$\boldsymbol{Z} = (\boldsymbol{z_1}\, \boldsymbol{z_2}\, \boldsymbol{z_3}) = \begin{bmatrix} 0,9777 & 0,4995 & -0,0084 \\ -0,8117 & 0,2802 & 0,0583 \\ 0,4315 & -0,7063 & 0,0250 \\ -0,5975 & -0,0734 & -0,0749 \end{bmatrix}.$$

Les moyennes des projections des individus sur chacun des axes sont nulles et leurs sommes de carrés sont égales aux valeurs propres et valent, respectivement, 2,1579 pour le premier axe, 0,8323 pour le deuxième axe et 0,0097 pour le troisième axe [3]. On en déduit que les différents axes contribuent à raison de 71,93 %, 27,75 % et 0,32 % à la reconstitution de $\boldsymbol{X}$ et que le plan factoriel défini par les deux premiers axes contribue à raison de 99,68 % à la reconstitution de $\boldsymbol{X}$.

Ces pourcentages donnent une idée globale de la qualité de la représentation du nuage des individus sur les axes ou le plan factoriel. Pour avoir une idée plus précise de la qualité de la représentation de chacun des individus sur les axes ou le plan factoriel, on peut calculer les ècosinus carrésè ou les ècontributions relatives des axes à la position des pointsè, définis au paragraphe 2.2. On trouve les valeurs données dans le tableau 1. On voit ainsi que le premier axe, qui intervient globalement pour 71,93 % dans la reconstitution de $\boldsymbol{X}$, intervient aussi pour 97,01 % dans la reconstitution du quatrième individu, mais uniquement pour 27,15 % pour la reconstitution du troisième individu. Par contre, on peut constater que les quatre points sont tous très bien représentés dans le plan factoriel 1-2.

Tableau 1. Qualité des représentations des individus sur chacun des trois axes factoriels et sur le premier plan factoriel : valeurs des cosinus carrés, en %.

Individus	Axe 1	Axe 2	Axe 3	Plan 1-2
1	79,29	20,70	0,01	99,99
2	88,94	5,51	0,46	99,54
3	27,15	72,75	0,09	99,91
4	97,01	1,46	1,52	98,48

D'autre part, on peut également calculer la part que prend chaque individu dans la somme des carrés associée à chacun des axes. Ainsi, pour le premier individu et le premier axe, on a :

$$0,9777^2/2,1579 = 0,4430 \text{ ou } 44,30\,\%\,.$$

Pour le même axe et pour les trois autres individus, on a respectivement 30,53 %, 8,63 % et 16,54 %. Des calculs analogues peuvent évidemment être réalisés pour les autres axes. Les valeurs ainsi obtenues sont appelées ècontributions des individus à la variance de l'axeè.

3. Dans les logiciels statistiques, les scores sont généralement multipliés par $\sqrt{n}$ ou $\sqrt{n-1}$, de sorte que leur variance et non leur somme de carrés soit égale à la valeur propre.

5.4. Représentation des variables

L'interprétation géométrique de la transformation de $\boldsymbol{Y}$ en $\boldsymbol{X}$ pour la représentation des variables dans l'espace des individus est différente de l'interprétation donnée pour la représentation des individus dans l'espace des variables. La soustraction de la moyenne de chacune des variables ne correspond plus à une translation de l'origine et la division par $s_j \sqrt{n}$ a comme effet de situer chaque point-variable à une distance unitaire de l'origine. En effet, pour chaque point-variable, le carré de la distance à l'origine d_{j0} vaut :

$$d_{j0}^2 = \sum_{i=1}^{n} x_{ij}^2 = \sum_{i=1}^{n} \left[(y_{ij} - \bar{y}_j) / (s_j \sqrt{n}) \right]^2 = 1 \,.$$

Le carré de la distance $d_{jj'}$ entre deux points-variables dans l'espace des individus est donné par la relation :

$$d_{jj'}^2 = \sum_{i=1}^{n} (x_{ij} - x_{ij'})^2 = \sum_{i=1}^{n} (x_{ij}^2 + x_{ij'}^2 - 2\, x_{ij}\, x_{ij'}) = 2\,(1 - r_{jj'})\,,$$

$r_{jj'}$ étant le coefficient de corrélation entre la variable j et la variable j'. Les distances entre les points-variables sont donc directement liées à la corrélation entre les variables : deux points sont très proches si leur corrélation est proche de 1 et ils sont d'autant plus éloignés que leur corrélation s'approche de -1.

Les vecteurs propres $\boldsymbol{v_1}$, $\boldsymbol{v_2}$ et $\boldsymbol{v_3}$ de la matrice $\boldsymbol{X}\,\boldsymbol{X'}$ sont les coefficients directeurs des trois premiers axes factoriels dans l'espace des individus, le quatrième axe étant sans intérêt puisque les projections des trois points-variables sur cet axe sont nulles. Les coordonnées des points sur les trois axes valent :

$$\boldsymbol{r_k} = \boldsymbol{X'}\,\boldsymbol{v_k} = \sqrt{\lambda_k}\;\boldsymbol{u_k}\,,$$

soit :

$$\boldsymbol{R} = (\boldsymbol{r_1}\,\boldsymbol{r_2}\,\boldsymbol{r_3}) = \begin{bmatrix} -0,5482 & 0,8363 & -0,0141 \\ 0,9402 & 0,3342 & 0,0665 \\ 0,9867 & -0,1462 & 0,0713 \end{bmatrix} .$$

Dans la mesure où les points-variables se trouvent sur une sphère de rayon unitaire, le carré de la coordonnée sur un axe donné est égal au cosinus carré défini au paragraphe 2.2. D'autre part, la coordonnée elle-même représente la corrélation qui existe entre la variable de départ, x_j ou y_j, et les valeurs des projections des points-individus sur cet axe. Ainsi, par exemple, la corrélation entre x_1 et z_1 est égale à :

$$\sum_{i=1}^{n} (x_{i1}\, z_{i1}) / \sqrt{\lambda_1}\,,$$

la somme des carrés des différentes colonnes de $\boldsymbol{X}$ étant égale à l'unité et la somme des carrés de z_1 étant égale à λ_1. Cette expression est bien égale à :

$$\boldsymbol{x_1'}\,\boldsymbol{z_1} / \sqrt{\lambda_1} = \boldsymbol{x_1'}\,\boldsymbol{v_1} = -\,0,5482\,,$$

$\boldsymbol{x_1'}$ étant la première ligne de $\boldsymbol{X}'$.

Les représentations graphiques des points-variables dans les plans factoriels sont de ce fait parfois appelés ècercles de corrélationè.

Les corrélations entre les variables initiales et les valeurs des projections des points-individus sur les axes factoriels jouent un rôle fondamental dans l'interprétation des axes. Pour l'exemple considéré, on voit que le premier axe, fortement corrélé avec le diamètre du lumen des fibres et l'épaisseur des fibres, est un axe de grosseur de fibres, tandis que le second axe est un axe de longueur des fibres.

Les carrés des coefficients de corrélation permettent de chiffrer la qualité des représentations des variables. Nous avons déjà vu que le premier axe intervient pour 71,93 % dans la reconstitution de $\boldsymbol{X}$. Sa contribution à la reconstitution de la première variable est égale à :

$$0,5482^2 = 0,3005 \text{ ou } 30,05\,\%\,,$$

et pour les deux autres variables, on a, respectivement, 88,40 % et 97,36 %. On voit également que le second axe est important pour la reconstitution de la première variable et, enfin, que les trois variables sont bien reconstituées à partir des deux axes factoriels : les pourcentages correspondant sont, en effet, égaux à 99,99, 99,57 et 99,50.

Dans la mesure où les colonnes de $\boldsymbol{X}$ contiennent les informations relatives à la variabilité des individus, la reconstitution d'une variable est une reconstitution de la variabilité des individus pour la variable en question. Par conséquent, les pourcentages de reconstitution des variables donnés ci-dessus correspondent en fait à des pourcentages de la variance des variables liée aux axes. Ainsi, par exemple, dire que le premier axe intervient pour 30 % dans la reconstitution de la première variable signifie aussi que 30 % de la variance de la variable en question est liée au premier axe.

6. ANALYSE DES CORRESPONDANCES

6.1. Transformation des données

Le point de départ de l'analyse des correspondances [4] est un tableau de fréquences à deux entrées. Ces fréquences constituent les éléments d'une matrice $\boldsymbol{Y}$, de dimensions $n \times p$, n et p représentant les nombres de modalités relatives aux deux critères pris en considération.

On constate donc que les lignes et les colonnes de la matrice $\boldsymbol{Y}$ sont de même nature, contrairement à la matrice des données pour une analyse en composantes principales, où les lignes correspondent aux individus et les colonnes aux variables.

Pour l'analyse des correspondances, la transformation appliquée à $\boldsymbol{Y}$ tient compte de l'identité de la nature des lignes et des colonnes, alors que pour

4. En anglais : *correspondence analysis*.

l'analyse des composantes, la transformation de $\boldsymbol{Y}$ en $\boldsymbol{X}$ a un effet différent sur les lignes et les colonnes. La matrice des fréquences $\boldsymbol{Y}$ est d'abord transformée en une matrice de fréquences relatives, $\boldsymbol{F}$, obtenue en divisant chaque élément de $\boldsymbol{Y}$ par la somme des éléments, c'est-à-dire par l'effectif total :

$$f_{ij} = y_{ij}/N \qquad (i = 1, \ldots, n\,;\; j = 1, \ldots, p)\,,$$

avec :

$$N = \sum_{i=1}^{n} \sum_{j=1}^{p} y_{ij}\,.$$

Cette matrice $\boldsymbol{F}$ est, à son tour, transformée en une matrice $\boldsymbol{X}$ par la transformation suivante :

$$x_{ij} = (f_{ij} - f_{i.}\, f_{.j})/\sqrt{f_{i.}\, f_{.j}}\,,$$

$f_{i.}$ et $f_{.j}$ étant les fréquences relatives marginales des lignes et des colonnes :

$$f_{i.} = \sum_{j=1}^{p} f_{ij} \quad \text{et} \quad f_{.j} = \sum_{i=1}^{n} f_{ij}\,.$$

On vérifie bien que, dans cette transformation, les lignes et les colonnes jouent un rôle symétrique. On peut remarquer aussi que les écarts :

$$f_{ij} - f_{i.}\, f_{.j}$$

sont en fait les écarts entre les fréquences relatives observées et les fréquences relatives attendues sous l'hypothèse d'indépendance entre les lignes et les colonnes. D'autre part, à la constante $1/\sqrt{N}$ près, les quantités x_{ij} sont égales aux racines carrées des quantités qu'on calcule pour chacune des np cellules du tableau de contingence et qu'on somme pour obtenir la valeur χ^2_{obs} lors de la réalisation du test χ^2 d'indépendance entre les lignes et les colonnes [Dagnelie, 1979-1980] :

$$\chi^2_{obs} = \sum_{i=1}^{n} \sum_{j=1}^{p} (N\, f_{ij} - N\, f_{i.}\, f_{.j})^2/(N\, f_{i.}\, f_{.j}) = N \sum_{i=1}^{n} \sum_{j=1}^{p} x_{ij}^2\,.$$

Il en résulte notamment que la trace des matrices $\boldsymbol{X}'\,\boldsymbol{X}$ et $\boldsymbol{X}\,\boldsymbol{X}'$, qui est égale à la somme des carrés de tous les éléments de $\boldsymbol{X}$, est égale à la valeur χ^2_{obs}, divisée par l'effectif total N.

A titre d'illustration, considérons les données reprises par Dagnelie [1979-1980], concernant l'influence de quatre traitements donnés sur la fructification du pommier *Golden Delicious*. Les lignes de la matrice $\boldsymbol{Y}$ correspondent aux quatre traitements (A, B, C et D) et les colonnes ont trait aux nombres de fruits produits par rameau (pas de fruit, un fruit, plus de un fruit) :

$$\boldsymbol{Y} = \begin{bmatrix} 203 & 150 & 6 \\ 266 & 112 & 1 \\ 258 & 126 & 2 \\ 196 & 168 & 17 \end{bmatrix}\,.$$

Figure 3. Calcul de la valeur χ^2_{obs} relative au test d'indépendance des lignes et des colonnes.

On en déduit :

$$\boldsymbol{F} = \begin{bmatrix} 0,1349 & 0,0997 & 0,0040 \\ 0,1767 & 0,0744 & 0,0007 \\ 0,1714 & 0,0837 & 0,0013 \\ 0,1302 & 0,1116 & 0,0113 \end{bmatrix}.$$

Les fréquences relatives marginales des lignes et des colonnes étant :

$$f_{1.} = 0,2385\,, \quad f_{2.} = 0,2518\,, \quad f_{3.} = 0,2565 \quad \text{et} \quad f_{4.} = 0,2532\,,$$

$$f_{.1} = 0,6133\,, \quad f_{.2} = 0,3694 \quad \text{et} \quad f_{.3} = 0,0173\,,$$

il en résulte que :

$$\boldsymbol{X} = \begin{bmatrix} -0,02983 & 0,03889 & -0,00209 \\ 0,05675 & -0,06103 & -0,05589 \\ 0,03564 & -0,03584 & -0,04660 \\ -0,06351 & 0,05920 & 0,10467 \end{bmatrix}.$$

La figure 3 reprend, pour l'exemple en question, un extrait du document imprimé obtenu par la procédure CORRESP du logiciel SAS. On y retrouve les données de départ, les fréquences attendues sous l'hypothèse d'indépendance entre les lignes et les colonnes, les écarts entre les fréquences observées et les fréquences attendues et, enfin, les contributions de chaque cellule du tableau à la valeur χ^2_{obs}. On vérifie bien que la racine carrée de ces éléments, divisée par la racine carrée de l'effectif total donne, au signe près, les éléments de la matrice $\boldsymbol{X}$. Par exemple, pour la première ligne et la première colonne, on a :

$$\sqrt{1,3391}/\sqrt{1.505} = 0,02983\,.$$

Le tableau des contributions à la valeur χ^2_{obs} donne également les totaux pour les lignes et les colonnes ainsi que le total général.

6.2. Valeurs propres, vecteurs propres et reconstitution du tableau de données

Du fait de la transformation utilisée, les matrices $\boldsymbol{X}'\boldsymbol{X}$ et $\boldsymbol{X}\boldsymbol{X}'$ sont toujours des matrices singulières et, pour l'exemple considéré, les deux valeurs propres non nulles sont égales à :

$$\lambda_1 = 0,03388 \quad \text{et} \quad \lambda_2 = 0,00181\,.$$

La somme des valeurs propres est égale à 0,03569 et on a bien :

$$\chi^2_{obs} = (0,03569)(1.505) = 53,71\,.$$

Les vecteurs propres correspondant aux deux valeurs propres non nulles sont, pour $\boldsymbol{X}'\,\boldsymbol{X}$ et pour $\boldsymbol{X}\,\boldsymbol{X}'$:

$$\boldsymbol{u_1} = \begin{bmatrix} -0,5212 \\ 0,5263 \\ 0,6718 \end{bmatrix} \quad \text{et} \quad \boldsymbol{u_2} = \begin{bmatrix} 0,3392 \\ -0,5946 \\ 0,7290 \end{bmatrix},$$

$$\boldsymbol{v_1} = \begin{bmatrix} 0,1880 \\ -0,5391 \\ -0,3734 \\ 0,7311 \end{bmatrix} \quad \text{et} \quad \boldsymbol{v_2} = \begin{bmatrix} -0,8170 \\ 0,3478 \\ -0,0135 \\ 0,4597 \end{bmatrix}.$$

La reconstitution de la matrice $\boldsymbol{X}$ par la somme des deux matrices :

$$\boldsymbol{X_k} = \sqrt{\lambda_k}\,\boldsymbol{v_k}\,\boldsymbol{u_k'} \qquad (k = 1,\, 2)\,,$$

conduit aux résultats suivants :

$$\boldsymbol{X} = \begin{bmatrix} -0,01804 & 0,01821 & 0,02325 \\ 0,05173 & -0,05223 & -0,06667 \\ 0,03583 & -0,03618 & -0,04618 \\ -0,07015 & 0,07083 & 0,09041 \end{bmatrix}$$

$$+ \begin{bmatrix} -0,01179 & 0,02067 & -0,02534 \\ 0,00502 & -0,00880 & 0,01079 \\ -0,00019 & 0,00034 & -0,00042 \\ 0,00663 & -0,01163 & 0,01426 \end{bmatrix}.$$

La première matrice, qu'on peut désigner par $\boldsymbol{X_1}$, est le produit de deux vecteurs qui minimise la somme des carrés des écarts par rapport aux éléments de $\boldsymbol{X}$. Les fréquences relatives reconstituées à partir de $\boldsymbol{X_1}$, et désignées par $f_{ij}^{(1)}$, sont telles que la quantité :

$$\sum_{i=1}^{n}\sum_{j=1}^{p}\left[\frac{(f_{ij}-f_{i.}\,f_{.j}}{\sqrt{f_{i.}\,f_{.j}}}-\frac{f_{ij}^{(1)}-f_{i.}\,f_{.j}}{\sqrt{f_{i.}\,f_{.j}}}\right]^2,$$

ou après simplification :

$$\sum_{i=1}^{n}\sum_{j=1}^{p}\frac{(f_{ij}-f_{ij}^{(1)})^2}{f_{i.}\,f_{.j}},$$

est minimum.

En multipliant par $\sqrt{f_{i.}\,f_{.j}}$ tous les éléments des matrices $\boldsymbol{X}$, $\boldsymbol{X_1}$ et $\boldsymbol{X_2}$, on obtient la formule de reconstitution des fréquences relatives, en termes d'écarts à la situation d'indépendance :

$$\boldsymbol{F}-\boldsymbol{F_0}=\boldsymbol{F_1}+\boldsymbol{F_2}\,,$$

l'élément ij de $\boldsymbol{F_0}$ étant égal à $f_{i.}\,f_{.j}$. En multipliant encore par l'effectif total, N, on a :

$$\boldsymbol{Y}-\boldsymbol{Y_0}=\boldsymbol{Y_1}+\boldsymbol{Y_2}\,,$$

avec :

$$\boldsymbol{Y_0}=\begin{bmatrix}220,17 & 132,63 & 6,20\\ 232,44 & 140,02 & 6,55\\ 236,73 & 142,60 & 6,67\\ 233,66 & 140,76 & 6,58\end{bmatrix},\quad \boldsymbol{Y_1}=\begin{bmatrix}-10,38 & 8,14 & 2,25\\ 30,59 & -23,98 & -6,62\\ 21,39 & -16,76 & -4,63\\ -41,60 & 32,60 & 9,00\end{bmatrix},$$

et

$$\boldsymbol{Y_2}=\begin{bmatrix}-6,79 & 9,24 & -2,45\\ 2,97 & -4,04 & 1,07\\ -0,12 & 0,16 & -0,04\\ 3,93 & -5,35 & 1,42\end{bmatrix}.$$

A la décomposition de la matrice $\boldsymbol{X}$ en une somme de deux matrices, $\boldsymbol{X_1}$ et $\boldsymbol{X_2}$, correspond aussi la décomposition de la valeur χ^2_{obs}. Il suffit pour cela d'élever au carré et de multiplier par N les éléments de $\boldsymbol{X_1}$ et de $\boldsymbol{X_2}$. Les résultats sont donnés dans les tableaux 2 et 3, de même que les totaux généraux. Les valeurs figurant dans ces tableaux peuvent également être obtenues en élevant au carré les éléments de $\boldsymbol{Y_1}$ ou de $\boldsymbol{Y_2}$ et en divisant les résultats par les éléments correspondants de $\boldsymbol{Y_0}$.

L'examen des tableaux 2 et 3 montre tout d'abord que le total des valeurs liées à $\boldsymbol{Y_1}$ et à $\boldsymbol{Y_2}$ est bien égal à la valeur χ^2_{obs} :

$$50,9950+2,7248=53,7199\,.$$

Tableau 2. Contribution au χ^2 de chaque cellule du tableau : valeurs pour le premier facteur.

i	j 1	2	3	Totaux
1	0,4897	0,4993	0,8136	1,8025
2	4,0270	4,1056	6,6902	14,8228
3	1,9321	1,9698	3,2099	7,1118
4	7,4053	7,5498	12,3028	27,2579
Totaux	13,8541	14,1245	23,0165	50,9950

Tableau 3. Contribution au χ^2 de chaque cellule du tableau : valeurs pour le deuxième facteur.

i	j 1	2	3	Totaux
1	0,2092	0,6431	0,9665	1,8188
2	0,0379	0,1166	0,1752	0,3297
3	0,0000	0,0002	0,0003	0,0005
4	0,0662	0,2036	0,3060	0,5758
Totaux	0,3134	0,9635	1,4479	2,7248

Si on exprime les valeurs en pour cent du total, on retrouve les valeurs propres de $\boldsymbol{X}'\boldsymbol{X}$ ou de $\boldsymbol{X}\boldsymbol{X}'$, exprimées en pour cent de la trace :

$$50,9950/53,7199 = 0,03388/(0,03388 + 0,00181) = 0,949 \text{ ou } 94,9\,\%$$

et $$2,7248/57,7199 = 0,00181/(0,03388 + 0,00181) = 0,051 \text{ ou } 5,1\,\%\,.$$

Ces valeurs signifient que le premier tableau, $\boldsymbol{Y_1}$, apporte 94,9 % de l'information et que le deuxième tableau, $\boldsymbol{Y_2}$, apporte 5,1 % de l'information. $\boldsymbol{Y_1}$ donne donc une très bonne approximation des écarts par rapport à l'indépendance et $\boldsymbol{Y_2}$ n'est qu'une correction mineure de cette approximation.

La figure 4 reprend les informations fournies par le logiciel SAS, en ce qui concerne les valeurs propres et la décomposition correspondante de χ^2_{obs}. Les valeurs singulières mentionnées dans cette figure sont les valeurs singulières de la matrice $\boldsymbol{X}$, c'est-à-dire aussi les racines carrées des valeurs propres de $\boldsymbol{X}'\boldsymbol{X}$ et de $\boldsymbol{X}\boldsymbol{X}'$. Ces valeurs propres sont également dénommées èinertiesè. Une justification de cette appellation sera donnée au paragraphe 4.3.

L'additivité des valeurs χ^2 ne se vérifie pas seulement pour le total général des tableaux 2 et 3, mais aussi pour les totaux par ligne et par colonne. Par

Figure 4. Valeurs propres et décomposition du χ^2 total.

Figure 5. Analyse des lignes.

exemple, on a, pour la première ligne :

$$1,8025 + 1,8188 = 3,6213\,,$$

et pour la première colonne :

$$13,8541 + 0,3134 = 14,1675\,.$$

Ces totaux par ligne et par colonne peuvent encore être exprimés en pour cent du χ^2 total et on obtient :

$$3,6213/53,7199 = 0,067 \quad \text{et} \quad 14,1675/53,7199 = 0,264\,.$$

Les valeurs ainsi trouvées pour les différentes lignes et pour les différentes colonnes sont appelées inerties des lignes et des colonnes et sont données dans les figures 5 et 6.

Les totaux par ligne et par colonne des tableaux 2 et 3 peuvent aussi être exprimés en pour cent des totaux par ligne et par colonne du tableau des contributions au χ^2_{obs} donnés à la figure 3. Pour la première ligne et la première colonne on a, respectivement :

$$1,8025/3,6213 = 0,498 \quad \text{et} \quad 1,8188/3,6213 = 0,502\,,$$

$$13,8541/14,1676 = 0,978 \quad \text{et} \quad 0,3134/14,1676 = 0,022\,.$$

La valeur 0,498 correspond à la contribution relative du premier facteur à la première ligne. Elle indique que la ligne est moins bien reconstituée par ce premier facteur, c'est-à-dire par $\boldsymbol{X_1}$, que ne le sont, en moyenne, les lignes et les colonnes, puisque le premier facteur exprime 94,9 % de χ^2_{obs}.

Comme nous le justifierons par la suite, les contributions des facteurs aux lignes et aux colonnes sont souvent appelées ècosinus carrésè. Elles sont données pour les lignes à la figure 5 et pour les colonnes à la figure 6. On notera aussi que les cosinus carrés peuvent s'additionner, permettant ainsi de déterminer la qualité de la représentation des lignes et des colonnes dans les sous-espaces constitués de plusieurs facteurs. Pour l'exemple considéré, cette addition ne présente pas d'intérêt, puisqu'on ne dispose que de deux facteurs. Le logiciel SAS donne la somme des cosinus carrés pour l'ensemble des facteurs retenus, sous la dénomination è*quality*è, ce qui peut être utile pour des tableaux présentant un plus grand nombre de lignes et de colonnes, alors qu'on ne s'intéresse qu'à un nombre limité de facteurs.

Figure 6. Analyse des colonnes.

Si, pour un facteur donné, on exprime la contribution d'une ligne ou d'une colonne en pour cent du total relatif au facteur donné, on obtient les contributions des lignes et des colonnes au facteur. Pour la première ligne et la première colonne du premier facteur on a respectivement :

$$1,8025/50,9950 = 0,035 \quad \text{et} \quad 13,8541/50,9950 = 0,272\,.$$

Toutes les contributions des lignes et des colonnes aux deux facteurs sont données dans les figures 5 et 6 sous les titres "*partial contributions*". Ces figures présentent également des tableaux reprenant ces contributions sous forme d'indices. Pour chaque ligne et chaque colonne, le numéro d'ordre du facteur pour lequel la ligne ou la colonne a la plus forte contribution est donné sous la dénomination "*best*" : on voit ainsi que les trois dernières lignes et la première colonne ont une contribution plus grande au premier facteur qu'au deuxième et qu'on a la situation opposée pour la première ligne et les deux dernières colonnes.

Pour chaque facteur, les colonnes "dim1" et "dim2" indiquent par un indice différent de zéro le ou les points (lignes ou colonnes) qui ont les contributions les plus fortes au facteur. La valeur de cet indice est le numéro d'ordre du facteur pour lequel la ligne ou la colonne a la plus forte contribution. Le nombre de points ainsi identifiés pour un facteur dépend de la contribution des points au facteur. De façon plus précise, on identifie les k points qui ont les contributions les plus fortes au facteur, de sorte que la somme des contributions de ces k points dépasse une valeur définie par l'utilisateur ou fixée, par défaut, à 0,8. Ainsi, par exemple, on peut constater que la deuxième et la quatrième ligne contribuent le plus au premier facteur : dans le tableau des codes, ces lignes sont représentées par le nombre 1, car elles ont une plus forte contribution au premier facteur qu'au second (code 1 dans la colonne "best"). La somme des contributions de la deuxième et de la quatrième lignes :

$$0,2907 + 0,5345 = 0,8252\,,$$

étant supérieure à 0,8, les autres lignes seront affectées du code 0. Quant aux contributions des colonnes au premier facteur, on constate qu'il faut prendre en considération les trois points, pour que la contribution totale dépasse 0,8 et les codes correspondant sont donc 1, 2 et 2, la contribution de la première colonne au premier facteur étant plus importante que celle au deuxième facteur, alors qu'on a la situation inverse pour les deux dernières colonnes. On peut évidemment déterminer, de manière similaire, les codes pour le deuxième facteur.

6.3. Représentation des points-lignes et des points-colonnes

Nous avons vu, au paragraphe 2.2, que les projections des points-lignes et des points-colonnes sur les axes sont données par les relations suivantes :

$$\boldsymbol{z_k} = \boldsymbol{X}\,\boldsymbol{u_k} = \sqrt{\lambda_k}\,\boldsymbol{v_k} \quad \text{et} \quad \boldsymbol{w_k} = \boldsymbol{X}'\,\boldsymbol{v_k} = \sqrt{\lambda_k}\,\boldsymbol{u_k}\,.$$

En appliquant ces relations aux données de l'exemple, on trouve :

$$\boldsymbol{Z} = (\boldsymbol{z_1}\, \boldsymbol{z_2}) = \begin{bmatrix} 0,0346 & -0,0347 \\ -0,0992 & 0,0148 \\ -0,0687 & -0,0006 \\ 0,1346 & 0,0196 \end{bmatrix}$$

et

$$\boldsymbol{W} = (\boldsymbol{w_1}\, \boldsymbol{w_2}) = \begin{bmatrix} -0,0959 & 0,0144 \\ 0,0969 & -0,0253 \\ 0,1237 & 0,0310 \end{bmatrix}.$$

Ces projections permettent de retrouver directement les contributions définies ci-dessus. En effet, la contribution relative d'un facteur à la position d'un point-ligne ou d'un point-colonne correspond au carré du cosinus de l'angle formé par le vecteur représentant un point projeté sur un facteur et le vecteur représentant ce même point dans l'espace complet. L'interprétation est tout à fait similaire à celle que nous avons donnée lors de l'analyse en composantes principales. Par exemple, pour la première ligne et pour la première colonne on a, respectivement :

$$0,0346^2/(0,0346^2 + 0,0347^2) = 0,498$$

et

$$0,0959^2/(0,0959^2 + 0,01443^2) = 0,978\,.$$

Quant aux contributions des lignes (ou des colonnes) aux différents facteurs, elles sont égales au rapport entre le carré de la projection d'un point-ligne (ou colonne) sur un facteur, à la somme des carrés des projections de tous les points-lignes (ou colonnes) sur ce même facteur. Cette somme de carrés étant égale à la valeur propre, les contributions sont égales au rapport entre le carré de la projection d'un point sur un facteur et la valeur propre de ce facteur. Ainsi, la contribution de la première ligne au premier facteur vaut :

$$0,0346^2/0,03388 = 0,035\,,$$

et la contribution de la première colonne au premier facteur vaut :

$$0,0959^2/0,03388 = 0,271\,.$$

Pour les représentations graphiques proprement dites, on modifie quelque peu les projections en fonction des fréquences relatives marginales : on divise, en effet, les éléments de la $i^{ième}$ ligne de la matrice $\boldsymbol{Z}$ donnée ci-dessus par $\sqrt{f_{i.}}$ et les éléments de la $j^{ième}$ ligne de $\boldsymbol{W}$ par $\sqrt{f_{.j}}$. On obtient :

$$\boldsymbol{Z^*} = \begin{bmatrix} 0,0709 & -0,0712 \\ -0,1978 & 0,0295 \\ -0,1357 & 0,0011 \\ 0,2675 & 0,0389 \end{bmatrix} \quad \text{et} \quad \boldsymbol{W^*} = \begin{bmatrix} -0,1225 & 0,0184 \\ 0,1594 & -0,0416 \\ 0,9409 & 0,2360 \end{bmatrix}.$$

On peut montrer qu'à la suite de cette modification, la moyenne des projections des points-lignes sur un axe donné, pondérées par les $f_{i.}$, est égale à zéro,

tout comme la moyenne des projections des points-colonnes sur un axe donné, pondérées par les $f_{.j}$, est égale à zéro. Pour les projections des points relatifs à la première ligne et à la première colonne et pour le premier axe, par exemple, on a, respectivement et aux erreurs d'arrondis près :

$$(0,0709)(0,2385) + \ldots + (0,2675)(0,2532) = 0$$

et $$(-0,1225)(0,6133) + (0,1594)(0,3694) + (0,9409)(0,0173) = 0\,.$$

D'autre part, la variance des projections des points-lignes ou des points-colonnes sur un axe donné, pondérées par les fréquences relatives marginales est égale à la valeur propre correspondant à cet axe. On a, par exemple :

$$(0,0709^2)(0,2385) + \ldots + (0,2675^2)(0,2532) = 0,0339$$

et $$(-0,1225^2)(0,6133) + (0,1594^2)(0,3694) + (0,9409^2)(0,0173) = 0,0339\,.$$

Les coordonnées des points-lignes et des points-colonnes permettent également, en tenant compte des masses respectives, de retrouver les contributions des points aux facteurs. Par exemple, la contribution de la première ligne au premier facteur est égale à :

$$(0,0709^2)(0,2385)/0,0339 = 0,0353\,.$$

Dans ces calculs, les fréquences relatives marginales jouent donc le rôle de poids et c'est ce qui justifie le nom de èmasseè qui leur est parfois donné. Ces masses, ainsi que les coordonnées des projections des points-lignes et des points-colonnes sur les deux axes sont données aux figures 5 et 6. Une représentation graphique simultanée des points-lignes et des points-colonnes dans le plan factoriel est donnée à la figure 7. Contrairement à l'analyse en composantes principales, où on effectue habituellement des graphiques séparés pour les individus et les variables, en analyse des correspondances, la représentation simultanée est la règle.

Figure 7. Représentation dans le plan factoriel des points-lignes (traitements A, B, C et D) et des points-colonnes (0 = pas de fruit, 1 = un fruit et + = plus de un fruit).

6.4. Interprétation des résultats

Comme pour l'analyse en composantes principales, les représentations graphiques constituent un outil important, mais pas unique, de l'interprétation des données de départ.

Dans ces graphiques, la proximité de deux points-lignes ou de deux points-colonnes traduit la similitude des èprofilsè, c'est-à-dire des distributions conditionnelles, relatifs à ces deux lignes ou à ces deux colonnes. Toutefois, comme

pour l'analyse en composantes principales, on ne perdra pas de vue que deux points (lignes ou colonnes) peuvent être proches dans un sous-espace (sur un axe ou sur un plan factoriel) sans nécessairement être proches dans l'espace complet, c'est-à-dire l'espace de l'ensemble des facteurs. On ne pourra donc conclure à la similitude des profils que si les points correspondants sont bien représentés, c'est-à-dire si la somme des cosinus carrés pour le sous-espace est suffisamment grande.

En pratique, on repère en premier lieu les points-lignes et les points-colonnes qui ont une forte contribution aux facteurs utilisés pour la représentation graphique et qui ont, en même temps, une qualité de représentation satisfaisante. Pour ces points, on examine alors les projections sur les axes et plus particulièrement le signe de ces projections, de manière à mettre en évidence les éventuelles conjonctions ou oppositions (points-lignes ou points-colonnes avec projection de même signe ou de signes contraires). L'examen des données initiales ou, plus exactement, la comparaison des profils des lignes et des colonnes au profil moyen des lignes et des colonnes permet également de vérifier ces conjonctions ou oppositions.

Pour l'exemple considéré, les profils des lignes et des colonnes sont donnés à la figure 8, tandis que les profils moyens des lignes et des colonnes correspondent aux masses des points-colonnes (figure 6) et aux masses des points-lignes (figure 5). L'examen de la figure 7 et de ces différents profils fait apparaître, pour le premier axe, une opposition entre les traitements A et D par rapport aux traitements B et C. Les traitements A et D sont caractérisés par des profils présentant des valeurs inférieures au profil moyen des lignes dans la première colonne et des valeurs supérieures ou égales au profil moyen pour les deux dernières colonnes. Par contre, les valeurs sont supérieures dans la première colonne et inférieures dans les deux autres colonnes aux valeurs du profil moyen des lignes pour la deuxième et la troisième ligne. Cette opposition est surtout marquée pour les traitements B et D, qui ont les contributions au facteur les plus importantes.

Figure 8. Profils des lignes et des colonnes.

Pour les colonnes, on note une conjonction, c'est-à-dire une attraction, des deux dernières colonnes et une opposition de la première colonne à ces deux dernières. La comparaison des trois profils-colonnes au profil-colonne moyen montre, en effet, que la première colonne est caractérisée par un déficit dans la première et la quatrième lignes et par un excès dans la deuxième et la troisième lignes, alors qu'on observe la situation inverse dans les deux autres colonnes. On vérifie aussi que c'est pour la troisième colonne que les écarts par rapport au profil

moyen sont les plus importants : cette colonne possède, en effet, la contribution la plus importante au premier facteur.

L'inertie liée au deuxième facteur étant très faible, nous ne tenterons pas de l'interpréter.

L'interprétation de la proximité d'un point-ligne et d'un point-colonne repose largement sur la relation existant entre les positions des points-lignes et des points-colonnes sur un axe donné. La coordonnée d'un point-ligne est, à une constante près, la moyenne arithmétique pondérée des coordonnées des différentes colonnes, les facteurs de pondération étant les fréquences conditionnelles pour la ligne en question et la constante étant l'inverse de la racine carrée de la valeur propre. Ainsi, pour l'exemple, la coordonnée de la première ligne sur le premier axe est égale, aux erreurs d'arrondis près, à :

$$[(-0,1225)(0,5655) + (0,1594)(0,4178) + (0,9409)(0,0167)] / \sqrt{0,03388} = 0,0709\,.$$

De la même manière, on peut déterminer la coordonnée d'un point-colonne à partir des coordonnées des points-lignes, des fréquences conditionnelles pour la colonne en question et de la valeur propre. Pour la première colonne et le premier axe, on a ainsi :

$$[(0,0709)(0,2199) + \ldots + (0,2675)(0,2124)] / \sqrt{0,03388} = -0,1225\,.$$

A une constante près, la coordonnée d'un point-ligne (ou d'un point-colonne) est égale à la moyenne des coordonnées des points-colonnes (ou des points-lignes), cette moyenne étant obtenue en introduisant des facteurs de pondération, fonction du profil de la ligne (ou de la colonne) : on dit que la coordonnée d'un point-ligne (ou d'un point-colonne) est un barycentre des coordonnées des points-colonnes (ou des points-lignes).

On comprend donc immédiatement que deux lignes (ou deux colonnes) qui ont le même profil doivent nécessairement coïncider dans les représentations graphiques. D'autre part, on perèoit également l'utilité de la constante $1/\sqrt{\lambda_k}$, dans les formules de passage des coordonnées des points-colonnes aux coordonnées des points-lignes et vice-versa. Cette constante est supérieure ou égale à l'unité. En effet, si elle était plus petite que l'unité, les points-lignes seraient nécessairement situés, sur un axe donné, entre les deux points-colonnes extrêmes, puisque les coordonnées des points-lignes seraient plus faibles, en valeurs absolues, que les moyennes pondérées des coordonnées des colonnes. Mais le même raisonnement s'appliquerait aux coordonnées des colonnes calculées à partir des coordonnées des lignes : les points-colonnes devraient donc aussi se situer entre les points-lignes extrêmes, ce qui, géométriquement est impossible. Il faut donc que la constante soit supérieure ou égale à l'unité, ce qui implique que les valeurs propres doivent être inférieures ou égales à l'unité.

La relation entre les positions des points-lignes et des points-colonnes montre que, si une ligne présente une fréquence conditionnelle élevée pour une colonne, le point-ligne sera attiré par le point-colonne correspondant. De même, si une colonne présente une fréquence conditionnelle élevée pour une ligne, le point-colonne sera attiré par le point-ligne correspondant.

L'interprétation de la proximité d'un point-ligne et d'un point-colonne reste cependant délicate du fait, d'une part, de la constante $1/\sqrt{\lambda_k}$, qui décale la position des points-lignes (ou colonnes) par rapport aux barycentres des colonnes (ou des lignes) et, d'autre part, parce que le barycentre, bien qu'étant éventuellement fortement influencé par un élément dont le poids est grand, peut en réalité se trouver accidentellement plus près d'un autre élément que de l'élément influent. En effet, le résultat d'une moyenne pondérée peut très bien être plus proche d'une valeur ayant un faible poids que de la valeur ayant le poids le plus important. En d'autres termes aussi, la proximité d'un point-ligne et d'un point-colonne ne signifie pas que le point-ligne (ou le point-colonne) a l'effet attractif le plus important sur le point-colonne (ou le point-ligne). L'interprétation de la proximité d'un point-ligne et d'un point-colonne doit, par conséquent, toujours s'appuyer sur l'examen des profils.

Pour l'exemple envisagé, la figure 7 révèle les proximités de la première colonne et des traitements B et C, d'une part, de la deuxième colonne et des traitements A et D, d'autre part. Dans le premier cas, l'attraction traduit l'excès de la proportion de rameaux sans fruit dans le cas des traitements B et C, et dans le second cas, elle traduit l'excès de la proportion de rameaux avec un fruit dans le cas des traitements A et D. On constate aussi que l'attraction du point relatif à la troisième colonne et à la quatrième ligne (excès de la proportion de rameaux présentant plus de un fruit dans le cas du traitement D) ne se traduit pas, sur le graphique, par une proximité des deux points en question.

La relation existant entre les positions des points-lignes et des points-colonnes sur un axe donné permet également de positionner sur les plans factoriels des points supplémentaires, relatifs à des lignes ou des colonnes qui ne sont pas prises en considération lors du calcul des axes factoriels. Pour illustrer cette notion, considérons, par exemple, le point correspondant à la modalité èprésence de fruitsè. Il s'agit donc d'une colonne supplémentaire dont le profil est constitué des quatre fréquences relatives conditionnelles suivantes, respectivement pour les traitements A, B, C et D :

$$0,2680, \quad 0,1942, \quad 0,2199 \quad \text{et} \quad 0,3179\,.$$

La coordonnée de ce point supplémentaire sur le premier axe vaut par conséquent :

$$[(0,0709)(0,2680) + (-0,1978)(0,1942) + (-0,1357)(0,2199) + (0,2675)(0,3179)]\,/\sqrt{0,03388} = 0,1943\,.$$

De façon analogue on trouve la coordonnée du point supplémentaire sur le deuxième axe :

$$\big[(-0,0712)(0,2680) + (0,0295)(0,1942) + (-0,0011)(0,2199) + (0,0389)(0,3179)/\sqrt{0,00181}\big] = -0,0292\,.$$

Pour ce point supplémentaire, les cosinus carrés relatifs aux deux axes sont respectivement :

$$0,1943^2/(0,1943^2 + 0,0292^2) = 0,978$$

et

$$0,0292^2/(0,1943^2 + 0,0292^2) = 0,022\,.$$

Géométriquement, le point supplémentaire se situe sur la droite joignant les points relatifs aux deux colonnes dont on a sommé les effectifs, c'est-à-dire les points représentés par les symboles "1" et "+" dans la figure 7. Plus précisément, il correspond au centre de gravité des deux points en question, la coordonnée sur chacun des axes étant la moyenne des coordonnées des deux points, pondérée par les masses des points. Comme le point è1è a une masse plus de vingt fois supérieure à celle du point "+", le point supplémentaire est beaucoup plus proche du point è1è que du point "+".

6.5. L'analyse des correspondances multiples

Au paragraphe 4.1, nous avons signalé que le point de départ de l'analyse des correspondances est un tableau de contingence à deux critères, présentant respectivement n et p modalités. La méthode peut cependant être étendue aux tableaux à plus de deux critères et permet, par exemple, de traiter les résultats d'enquêtes au cours desquelles des personnes doivent répondre à un ensemble de questions à choix multiples. On parle alors d'analyse des correspondances multiples, qui consiste à réaliser une analyse des correspondances sur un tableau résultant d'un codage particulier des réponses aux questions et qu'on appelle tableau disjonctif complet. Un tel tableau comporte une ligne par personne interrogée et autant de colonnes qu'il y a de modalités de réponses pour l'ensemble des questions. A l'intérieur de ce tableau on indique, pour chaque personne interrogée et pour chaque question la valeur 1 en regard de la modalité qui correspond à la réponse de la personne et la valeur 0 en regard des modalités qui ne correspondent pas à la réponse de la personne. Bien sđr, la méthode s'applique également à d'autres situations que les enquêtes et peut être utilisée aussi en présence de variables quantitatives, celles-ci étant remplacées par un nombre fini de modalités différentes.

Afin d'illustrer la notion de tableau disjonctif complet, nous avons repris l'exemple relatif à l'expérience sur les pommiers, bien qu'il n'y ait que deux critères. Le tableau 4 donne ce tableau disjonctif complet, mais uniquement pour un individu par catégorie. Pour obtenir le tableau complet de 1.505 lignes, il suffirait de répéter 203 fois la première ligne, 266 fois la seconde ligne, et ainsi de suite. En fait, l'analyse de ce tableau de 1.505 lignes serait tout à fait équivalente à l'analyse du tableau de 12 lignes (tableau 4) dans lequel les deux valeurs unitaires présentes sur chaque ligne seraient remplacées par la fréquence des lignes identiques à la ligne en question. En analyse des correspondances, on peut, en effet, toujours additionner des lignes ou des colonnes ayant le même profil, sans altérer les résultats. Cette propriété est connue sous la dénomination de principe d'équivalence distributionnelle.

Si on désigne par $\boldsymbol{Y}$ le tableau disjonctif complet, le produit $\boldsymbol{Y}'\boldsymbol{Y}$ donne naissance à un tableau, appelé tableau de Burt, qui croise les unes avec les autres toutes les modalités de chacun des critères. Dans le cas particulier de deux critères, ce tableau de Burt n'est rien d'autre que le tableau de contingence classique.

L'interprétation des résultats de l'analyse des correspondances d'un ta-

Tableau 4. Tableau disjonctif complet relatif aux deux critères de classification, à raison d'un seul individu par catégorie.

Catégories	Traitements				Nombres de fruits		
	A	B	C	D	0	1	+
a	1	0	0	0	1	0	0
b	0	1	0	0	1	0	0
c	0	0	1	0	1	0	0
d	0	0	0	1	1	0	0
e	1	0	0	0	0	1	0
f	0	1	0	0	0	1	0
g	0	0	1	0	0	1	0
h	0	0	0	1	0	1	0
i	1	0	0	0	0	0	1
j	0	1	0	0	0	0	1
k	0	0	1	0	0	0	1
l	0	0	0	1	0	0	1

Figure 9. Valeurs propres et décomposition du χ^2 total pour le tableau disjonctif complet.

bleau disjonctif complet doit tenir compte du caractère spécifique de ce type de tableau. En particulier, la somme des valeurs propres fournies par l'analyse des correspondances dépend du nombre de critères et du nombre de modalités pour l'ensemble des critères, et non des résultats relatifs aux différents individus. Pour q critères et, au total, p modalités, cette somme est égale à $(p-q)/q$ et le nombre maximum de valeurs propres non nulles est égal à $p-q$. D'autre part, contrairement à l'analyse des correspondances simples, la situation d'indépendance se traduit, non pas par la nullité des valeurs propres, mais par $p-q$ valeurs propres identiques et égales à $1/q$.

Nous allons illustrer l'analyse des correspondances multiples en traitant les données relatives à l'expérience sur pommiers à partir du tableau disjonctif complet, bien qu'il n'y ait que deux critères de classification. La figure 9 donne les valeurs propres et la décomposition du χ^2 et la figure 10 reprend le graphique des points-lignes et des points-colonnes dans le plan des deux premiers facteurs. On vérifie bien que le nombre de valeurs propres est égal à 5 et que la somme des valeurs propres est égale à 2,5, puisqu'on dispose d'un total de cinq modalités pour les deux critères pris en considération.

On constate aussi une nette similitude entre la position des points-colonnes correspondant aux sept modalités dans la figure 10 et les points-lignes et les points-colonnes de la figure 7 : il suffirait de multiplier par 0,2392 les abscisses et par 0,0589 les ordonnées des points de la figure 10 pour obtenir les coordonnées

Figure 10. Représentation des points-lignes et des points-colonnes dans le premier plan factoriel : les points-colonnes sont représentés par les symboles donnés à la figure 7 (A, B, C, D, 0, 1 et +) et les points-lignes sont représentés par les symboles donnés dans le tableau 4 (a, b, ..., l).

des points de la figure 8. Ces constantes sont en fait les racines carrées des rapports des valeurs propres relatives aux deux analyses :

$$\sqrt{0,03388/0,59204} = 0,2392 \quad \text{et} \quad \sqrt{0,00181/0,52128} = 0,0589\,.$$

D'autre part, les positions des 12 points-lignes, qui correspondent aux 12 catégories de rameaux de pommiers résultant du croisement des deux critères de classification, montrent nettement l'attirance des rameaux avec plus d'un fruit (points i, j, k et l) avec le point relatif à la modalité èplus d'un fruitè et avec le point relatif au traitement D. De même, les points a, b, c et d sont dans l'ensemble relativement proches de la modalité èpas de fruitè, et les points e, f, g et h sont relativement proches de la modalité èun fruitè. D'autre part, on remarque aussi que les quatre points-lignes relatifs à chacun des quatre traitements ont, les uns par rapport aux autres, une position relative assez stable, quels que soient les nombres de fruits considérés, et ces positions relatives sont proches des positions relatives des points-colonnes correspondant aux quatre traitements eux-mêmes (points A, B, C et D).

7. QUELQUES INFORMATIONS COMPLÉMENTAIRES

Au cours des paragraphes précédents, nous nous sommes efforcé d'expliquer la succession des calculs effectués lors d'une analyse en composantes principales et d'une analyse des correspondances, en insistant particulièrement sur les éléments utiles lors de l'interprétation des résultats. Dans ce but, nous avons envisagé deux exemples numériques de dimensions réduites, qui ne révèlent sans doute pas les possibilités réelles des méthodes décrites. D'autres illustrations et des informations relatives à l'interprétation des résultats peuvent être trouvées dans les références données dans l'introduction (paragraphe 1), ainsi que dans les documents de Dervin [1988] et de Philippeau [1986].

Nous nous sommes également limité à l'étude de deux cas particuliers et nous n'avons pas envisagé l'analyse factorielle classique qui, elle aussi, dérive des principes généraux exposés au paragraphe 2. Des informations au sujet de ce type d'analyse peuvent être trouvées, notamment, dans Dagnelie [1982].

De plus, en relation avec l'analyse en composantes principales, on peut encore mentionner l'analyse des matrices de corrélation de rang et l'analyse des matrices de corrélation (ou de variances et covariances) partielle. Il s'agit, en fait, d'analyses en composantes principales dans lesquelles les coefficients de corrélation classiques repris dans la matrice $\boldsymbol{X}'\boldsymbol{X}$ sont remplacés par des coefficients

de corrélation de rang de SPERMAN ou par des coefficients de corrélation partielle (ou des covariances partielles). Des informations à ce sujet sont données par LEBART *et al.* [1979].

On peut constater que, en fin de compte, les méthodes factorielles sont assez variées et le choix d'une méthode particulière peut prêter à discussion. Des éléments de comparaison de ces méthodes sont donnés par CIBOIS [1983].

D'autre part, pour présenter l'analyse en composantes principales et l'analyse des correspondances comme des applications particulières de l'analyse générale, nous avons, dans un premier temps, transformé la matrice des données initiales, $\boldsymbol{Y}$, en une matrice $\boldsymbol{X}$.

Pour l'analyse en composantes principales, nous avons utilisé une transformation telle que la matrice $\boldsymbol{X}'\boldsymbol{X}$ soit la matrice de corrélation, mais nous avons signalé également que, pour certains problèmes, il pouvait se justifier d'effectuer les calculs à partir de la matrice des variances et covariances (paragraphe 2.1). Bien entendu, l'analyse réalisée à partir de la matrice des variances et covariances donne, d'une faèon générale, des résultats différents de ceux obtenus à partir de la matrice de corrélation.

Pour l'analyse des correspondances, nous avons pris comme point de départ, la matrice $\boldsymbol{X}$, dont les éléments sont fonction des écarts par rapport à la situation d'indépendance entre les lignes et les colonnes (paragraphe 4.1). On aurait cependant pu faire également l'analyse à partir d'une matrice $\boldsymbol{X}^*$ dont les éléments seraient égaux à :

$$\boldsymbol{x}^*_{\boldsymbol{ij}} = \frac{f_{ij}}{\sqrt{f_{i.} f_{.j}}} \,.$$

Dans ce cas, la première valeur propre de $\boldsymbol{X}^{*\prime}\boldsymbol{X}^*$ aurait été égale à l'unité et les autres valeurs propres auraient été identiques aux valeurs propres de la matrice $\boldsymbol{X}'\boldsymbol{X}$ définie au paragraphe 4.1. La première valeur propre, toujours égale à l'unité pour ce type de transformation, est dite triviale et est écartée. Elle permet, en fait, de donner comme première approximation du tableau des fréquences initiales, la matrice $\boldsymbol{Y_0}$ définie au paragraphe 4.2. L'existence de cette valeur propre triviale, qui n'apparaît en général pas explicitement dans les documents de sortie des programmes d'analyse des correspondances ne doit cependant pas être perdue de vue lorsqu'on souhaite comparer les résultats d'une analyse des correspondances à ceux d'une analyse en composantes principales effectuée sur les mêmes données : du fait de l'élimination de la valeur propre triviale, la composante k de l'analyse des correspondances ne doit pas être comparée à la composante k de l'analyse en composantes principales mais bien à la composante $k+1$ [CIBOIS, 1983].

Par ailleurs, l'analyse des correspondances, du moins dans le cas des tableaux de contingence à deux entrées, peut aussi être considérée comme une application particulière de l'analyse des corrélations et des variables canoniques, dont une description est donnée notamment par DAGNELIE [1982] et PALM [1990].

Il suffit de soumettre à l'analyse canonique le tableau disjonctif complet, les p colonnes relatives aux p modalités du premier critère constituant les variables du premier groupe et les q colonnes relatives aux q modalités du second critère constituant les variables du second groupe. On peut calculer $p-1$ (ou $q-1$, si $q<p$) coefficients de corrélation canonique qui correspondent aux racines carrées des valeurs propres de l'analyse du tableau de contingence de dimensions $p \times q$, et les graphiques des individus dans les plans des variables canoniques de chacun des deux groupes correspondent, à un facteur d'échelle près, aux représentations des points-lignes ou des points-colonnes dans les espaces des facteurs.

A titre d'illustration, le tableau disjonctif complet relatif à l'expérience menée sur les pommiers (tableau 4) a été soumis à une analyse des corrélations canoniques, en tenant compte, bien sđr, des fréquences d'observation de chacune

des lignes de ce tableau. On a obtenu deux coefficients de corrélation canonique différents de zéro :

$$r_1 = 0,184075 \quad \text{et} \quad r_2 = 0,042550\,.$$

Elevées au carré, ces valeurs sont bien égales aux valeurs propres qui apparaissent dans la figure 2.

Les figures 11 et 12 donnent les représentations des individus dans le plan des deux variables canoniques relatives aux variables du groupe ètraitementsè et dans le plan des deux variables canoniques relatives aux variables du groupe "nombres de fruits". Le premier graphique contient quatre points, qui correspondent aux quatre traitements. Il s'agit en fait de points multiples, car les 1.505 rameaux se répartissent en quatre catégories en ce qui concerne les variables du groupe ètraitementsè, et les individus a, b, c et d du tableau 4 sont représentatifs de ces catégories. De même, le second graphique contient trois points multiples correspondant aux trois classes de nombres de fruits, l'individu a étant un représentant des rameaux sans fruit, l'individu c étant un représentant des rameaux avec un fruit et l'individu i étant un représentant des rameaux à plus d'un fruit.

Figure 11. Représentation des quatre traitements dans le plan des deux premières variables canoniques relatives aux variables du groupe "traitement" (a = traitement A, b = traitement B, c = traitement C et d = traitement D).

Les variables canoniques étant, par construction, de variance unitaire, alors que les projections des points sur un axe en analyse des correspondances sont de variance égale à la valeur propre, il faut multiplier les coordonnées des points sur les deux premiers axes canoniques par les racines carrées des valeurs propres pour retrouver les projections sur les axes factoriels. Ainsi, pour le point a, par exemple, les coordonnées sur les axes canoniques valent $-0,38482$ et 1,67225. Les coordonnées dans le plan factoriel sont donc :

$$(-0,38482)\sqrt{0,03388} = -0,07083 \quad \text{et} \quad (1,67225)\sqrt{0,00181} = 0,07114\,.$$

On retrouve bien, aux erreurs d'arrondis et au signe près, qui est arbitraire, les coordonnées du point-ligne relatif au premier traitement, données dans la figure 5. Si on effectuait des calculs identiques pour les six autres points des figures 11 et 12, et si on reportait l'ensemble des points sur un même graphique, on retrouverait la figure 7.

Nous venons d'illustrer le lien qui existe entre l'analyse canonique et l'analyse des correspondances simples. En fait, on peut montrer aussi que l'analyse des correspondances multiples est un cas particulier de l'analyse canonique généralisée à plus de deux groupes de variables, pour laquelle des informations sont données, notamment, par Harris [1975], Morrisson [1967] et Press [1972].

Figure 12. Représentation des trois modalités de nombres de fruits dans les plans des deux premières variables canoniques relatives aux variables du groupe "nombres de fruits" (a = pas de fruit, c = un fruit, i = plus de un fruit).

Enfin pour terminer, deux remarques peuvent encore être faites à propos des représentations graphiques des points dans les différents plans factoriels.

La première remarque a trait aux échelles utilisées. Lorsque ces graphiques sont réalisés par des procédures standards d'établissement de diagrammes de dispersion, les axes sont souvent gradués de manière automatique, afin d'assurer aux graphiques une forme relativement carrée. Il en résulte qu'une unité sur un axe n'a en général pas la même longueur qu'une unité sur l'autre axe. Dans ce cas, les distances réelles entre les points ne sont pas proportionnelles aux distances qu'on observe sur le graphique, ce qui peut conduire à des erreurs d'interprétation. Dans la mesure du possible, on veillera donc à respecter les échelles sur les axes afin de visualiser les différences de dispersion selon les deux axes en question.

La seconde remarque est spécifique à l'analyse en composantes principales. Nous avons vu, aux paragraphes 3.3 et 3.4, que les individus et les variables, c'est-à-dire les lignes et les colonnes, font en général l'objet de représentations graphiques séparées, contrairement à l'analyse des correspondances. Il faut noter à ce sujet que l'examen détaillé du graphique des individus ne se justifie que si les individus présentent un intérêt par eux-mêmes. Ce ne sera, par exemple, habituellement pas le cas si ceux-ci sont prélevés d'une manière aléatoire et simple dans une population. La représentation graphique permet alors essentiellement la visualisation d'éventuels individus anormaux. Lorsque la représentation des individus présente un intérêt, et si on souhaite superposer la représentation des individus et des variables, on peut multiplier les coordonnées des individus par une constante, de manière à ce que la distance moyenne d'un individu au centre de gravité soit égale à l'unité, comme la distance de chacune des variables à ce centre de gravité [Lebart *et al.*, 1979]. Si les coordonnées des individus sont telles que la variance observée pour chacun des axes est égale à la valeur propre, comme c'est habituellement le cas, on multiplie les coordonnées des individus par $1/\sqrt{p}$.

8. BIBLIOGRAPHIE

Benzécri J.P., Benzécri F. [1980]. *Pratique de l'analyse des données. 1. Analyse des correspondances : exposé élémentaire.* Paris, Dunod, 424 p.

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

Cibois P. [1983]. *L'analyse factorielle.* Paris, Presses universitaires de France, 128 p.

Dagnelie P. [1979-1980]. *Théorie et méthodes statistiques : applications agronomiques* (2 vol.). Gembloux, Presses agronomiques, 378 + 463 p.

Dagnelie P. [1982]. *Analyse statistique à plusieurs variables*. Gembloux, Presses agronomiques, 362 p.

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

Fénelon J.P. [1981]. *Qu'est-ce que l'analyse des données* ? Paris, Lefonen, 311 p.

Greenacre M.J. [1984]. *Theory and application of correspondence analysis*. London, Academic Press, 364 p.

Harris R. [1975]. *A primer of multivariate statistics*. New York, Academic Press, 332 p.

Jackson J.E. [1991]. *A user's guide to principal components*. New York, Wiley, 570 p.

Lebart L., Morineau A., Fénelon J.P. [1979]. *Traitement des données statistiques - méthodes et programmes*. Paris, Dunod, 510 p.

Leclercq A. [1979]. *Influence du milieu et du traitement sur la qualité du bois de hêtre*. Gembloux, Faculté des Sciences agronomiques, 339 p.

Morrisson D. [1967]. *Multivariate statistical methods*. New York, McGraw Hill, 338 p.

Palm R. [1990]. La corrélation canonique : principes et application. *Notes Stat. Inform.* (Gembloux), 90/1, 28 p.

Philippeau G. [1986]. *Comment interpréter les résultats d'une analyse en composantes principales ?* Paris, Institut technique des Céréales et des Fourrages, 63 p.

Press J. [1972]. *Applied multivariate analysis*. New York, Holt, Rinehart and Winston, 521 p.

X [1990]. *SAS/STAT user's guide, version 6* (vol. 1). Cary, SAS Institute, 889 + 53 p.

La collection

NOTES DE STATISTIQUE ET D'INFORMATIQUE

réunit divers travaux (notes techniques, rapports de recherche, publications, etc.) émanant des services de statistique et d'informatique de la Faculté des Sciences Agronomiques et du Centre de Recherches Agronomiques de Gembloux (Belgique).

Quelques titres récents:

RAMLOT P. [1986]. Programmation structurée en Cobol, en Fortran et en Basic: aspects méthodologiques. *Notes Stat. Inform.* (Gembloux) 86/7, 33 p.

RAMLOT P., TOUSSAINT A., VANDEVANDEL J.P. [1986]. Programmation structurée en Cobol, en Fortran et en Basic: étude de cas. *Notes Stat. Inform.* (Gembloux) 86/8, 36 p.

PALM R. [1987]. Etude des séries chronologiques par les méthodes de décomposition. *Notes Stat. Inform.* (Gembloux) 97/1, 25 p.

PALM R. [1987]. Etude des séries chronologiques par la méthode de BOX et JENKINS. *Notes Stat. Inform.* (Gembloux) 87/2, 40 p.

PALM [1988]. Les critères de validation des équations de régression linéaire. *Notes Stat. Inform.* (Gembloux) 88/1, 27 p.

PALM R., DE BAST A; [1988]. Construction d'un modèle agrométéorologique pour la prévision des productions agricoles dans la Communauté Economique Européenne. *Notes Stat. Inform.* (Gembloux) 88/2, 14 p.

PALM [1989]. Quelques éléments de programmation linéaire *Notes Stat. Inform.* (Gembloux) 89/1, 37 p.

DAGNELIE P. [1989]. Le choix d'une méthode d'analyse statistique et l'examen préliminaire des données. *Notes Stat. Inform.* (Gembloux) 89/2, 17 p.

PALM R. [1990]. La corrélation canonique: principes et application. *Notes Stat. Inform.* (Gembloux) 90/1, 28 p.

CLAUSTRIAUX J.J. [1990]. Introduction au logiciel statistique MINITAB. *Notes Stat. Inform.* (Gembloux) 90/2, 12 p.

CARLETTI G. [1991]. Les micro-ordinaeurs: présentation générale. *Notes Stat. Inform.* (Gembloux) 91/1, 27 p.

IEMMA A.F., PALM R. [1992]. Les matrices généralisées et leur utilisation dans le modèle linéaire. *Notes Stat. Inform.* (Gembloux) 92/1, 25 p.

Faculté universitaire des Sciences agronomiques
Avenue de la Faculté d'Agronomie 8
5030 GEMBLOUX (Belgique)

D/1993/2371/1