NOTES DE STATISTIQUE
ET D'INFORMATIQUE

2012/2

LA REPRÉSENTATION D'UNE MATRICE PAR BIPLOT

R. PALM, C. CHARLES et J. J. CLAUSTRIAUX

Université de Liège – Gembloux Agro-Bio Tech
Unité de Statistique, Informatique et Mathématique appliquées à la bioingénierie
GEMBLOUX
(Belgique)

LA REPRÉSENTATION D'UNE MATRICE PAR BIPLOT

R. Palm [*], C. Charles [†] et J. J. Claustriaux [‡]

RÉSUMÉ

La factorisation d'une matrice par décomposition par valeurs singulières est examinée et utilisée pour la représentation d'une matrice par biplot. Le lien entre cette représentation et les représentations habituelles des variables et des individus réalisées lors des analyses en composantes principales est ensuite examiné et illustré par un exemple numérique.

SUMMARY

Matrix factorization by means of singular value decomposition is examined and used to produce a graphical representation of a data matrix called biplot. The link between this biplot and the plots of the variables and of the individuals usually given in principal component analysis is discussed and applied to an example.

1. INTRODUCTION

Un *biplot* est une représentation graphique d'une matrice, généralement dans un espace à deux dimensions, chaque ligne et chaque colonne étant représentée par un point.

Le préfixe *bi* fait allusion aux deux types de points et non à la dimension de l'espace servant à la représentation. D'un point de vue théorique, la dimension de l'espace peut être augmentée sans difficulté, mais au-delà de la dimension 3, la visualisation ne peut se faire que par des projections des points dans des sous-espaces.

La matrice qui fait l'objet de la représentation n'est généralement pas la matrice des observations brutes. Les variables sont en effet le plus souvent centrées, par soustraction de la moyenne, et standardisées d'une manière ou d'une autre.

[*]Professeur à l'Université de Liège, Gembloux Agro-Bio Tech.
[†]Chargée de cours à l'Université de Liège, Gembloux Agro-Bio Tech.
[‡]Professeur ordinaire à l'Université de Liège, Gembloux Agro-Bio Tech.

Le biplot repose sur la factorisation des matrices. La technique générale est présentée au paragraphe 2 et une factorisation particulière, la décomposition par valeurs propres, est décrite au paragraphe 3. Les propriétés de trois formes de biplot, basées sur cette décomposition par valeurs singulières, sont examinées au paragraphe 4. La relation entre les biplots et l'analyse en composantes principales fait l'objet du paragraphe 5. Enfin, quelques informations complémentaires sont données au paragraphe 6.

La représentation par biplot a été proposée par Gabriel [1971]. Le lecteur y trouvera une description approfondie de la méthode. Des informations peuvent également être trouvées dans plusieurs livres consacrés à l'analyse statistique multivariée, en particulier dans les ouvrages de Jackson [1991], Khattree et Naik [1995], Rencher [2002] et Seber [1984], ainsi que dans les monographies consacrées entièrement aux biplots de Gower et Hand [1996] et de Gower *et al.* [2011].

2. FACTORISATION, POINTS-LIGNES ET POINTS-COLONNES

2.1. Factorisation d'une matrice

Soit $\boldsymbol{Y}$ une matrice de dimensions $n \times p$ et de rang r. Il pourrait s'agir, par exemple, d'une matrice d'observations réalisées sur n individus pour p variables. Cette matrice $\boldsymbol{Y}$ peut être exprimée sous la forme d'un produit de deux matrices :

$$\boldsymbol{Y} = \boldsymbol{A}\,\boldsymbol{B}',$$

$\boldsymbol{A}$ étant une matrice de dimensions $n \times r$ et de rang r et $\boldsymbol{B}$ une matrice de dimensions $p \times r$ et de rang r également. Ces matrices $\boldsymbol{A}$ et $\boldsymbol{B}$ qui interviennent dans la factorisation de $\boldsymbol{Y}$ ne sont pas uniques et nous examinerons, au paragraphe 3, comment elles peuvent être déterminées.

Chaque élément y_{ij} de $\boldsymbol{Y}$ est égal au produit de deux vecteurs :

$$y_{ij} = \boldsymbol{a}_i\,\boldsymbol{b}'_j \qquad (i = 1, ..., n;\ j = 1, ..., p),$$

$\boldsymbol{a}_i$ étant la $i^{\text{ème}}$ ligne de $\boldsymbol{A}$ et $\boldsymbol{b}_j$ la $j^{\text{ème}}$ ligne de $\boldsymbol{B}$.

Cette relation peut encore s'écrire :

$$y_{ij} = a_{i1}\,b_{j1} + a_{i2}\,b_{j2} + ... + a_{ir}\,b_{jr}.$$

A titre d'illustration, considérons la matrice quelconque suivante :

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

qui peut être factorisée par les matrices suivantes :

$$\boldsymbol{A} = \begin{bmatrix} 3,11 & -1,31 & 0,21 \\ 1,56 & -1,37 & -0,05 \\ 2,31 & 0,10 & -0,63 \\ 4,46 & 1,34 & 0,20 \end{bmatrix} \text{ et } \quad \boldsymbol{B} = \begin{bmatrix} 4,47 & -0,76 & 0,41 \\ 2,50 & 2,12 & 0,00 \\ 3,33 & -0,57 & -0,56 \end{bmatrix}.$$

On notera que les éléments des matrices $\boldsymbol{A}$ et $\boldsymbol{B}$ ont été arrondis à deux décimales. Les calculs ultérieurs ont toutefois été réalisés avec un plus grand nombre de décimales. Le lecteur qui referait les calculs de façon manuelle pourrait donc obtenir des résultats légèrement différents de ceux reproduits ici.

On peut vérifier que, par exemple, l'élément y_{32} est bien égal aux produits des vecteurs $\boldsymbol{a}_3$ et $\boldsymbol{b}'_2$, avec :

$$\boldsymbol{a}_3 = (2,31 \quad 0,10 \quad -0,63) \quad \text{et} \quad \boldsymbol{b}_2 = (2,50 \quad 2,12 \quad 0,00).$$

En effet :

$$y_{32} = (2,31)(2,50) + (0,10)(2,12) - (0,63)(0,00) \simeq 6.$$

Des calculs similaires pourraient être réalisés pour les douze éléments de $\boldsymbol{Y}$.

2.2. Points-lignes et points-colonnes pour une matrice de rang deux

Nous considérons d'abord la représentation graphique d'une matrice de rang deux. Les matrices $\boldsymbol{A}$ et $\boldsymbol{B}$ n'ont alors que deux colonnes et les vecteurs $\boldsymbol{a}_i$ et $\boldsymbol{b}_j$ n'ont que deux éléments. Les n vecteurs $\boldsymbol{a}_i$ et les p vecteurs $\boldsymbol{b}_j$ peuvent être représentés dans un espace de dimension 2. Et pour distinguer les vecteurs $\boldsymbol{a}_i$ des vecteurs $\boldsymbol{b}_j$ dans ce graphique, on représente une série de vecteurs, par exemple les $\boldsymbol{a}_i$, par des points situant l'extrémité des vecteurs et l'autre série de vecteurs, les $\boldsymbol{b}_j$, par des flèches rejoignant l'origine des axes aux extrémités des vecteurs. Ces vecteurs seront appelés par la suite, respectivement les points-lignes et les points-colonnes, car les lignes de $\boldsymbol{A}$ sont associées aux lignes de $\boldsymbol{Y}$ et les lignes de $\boldsymbol{B}$ sont associées aux colonnes de $\boldsymbol{Y}$.

Les vecteurs $\boldsymbol{a}_i$ et $\boldsymbol{b}_j$ présentent trois propriétés qui seront utilisées par la suite et que nous énonçons ci-dessous.

1. La longueur d'un vecteur (longueur du segment rejoignant l'origine au point pour un vecteur $\boldsymbol{a}_i$ ou longueur de la flèche pour un vecteur $\boldsymbol{b}_j$), appelée aussi norme, est égale à la racine carrée du produit du vecteur par sa transposée :

$$\|\boldsymbol{a_i}\| = (\boldsymbol{a}_i \, \boldsymbol{a}'_i)^{1/2} = \sqrt{a_{i1}^2 + a_{i2}^2}$$

et

$$\|\boldsymbol{b_j}\| = (\boldsymbol{b}_j \, \boldsymbol{b}'_j)^{1/2} = \sqrt{b_{j1}^2 + b_{j2}^2}.$$

2. Le produit scalaire (ou produit interne) des vecteurs $\boldsymbol{a}_i$ et $\boldsymbol{b}_j$ est égal à y_{ij} :

$$< \boldsymbol{a_i}, \boldsymbol{b_j} >= \|\boldsymbol{a_i}\| \, \|\boldsymbol{b_j}\| \cos\theta_{ij} = a_{i1} \, b_{j1} + a_{i2} \, b_{j2} = y_{ij},$$

θ_{ij} étant l'angle formé par les deux vecteurs. L'observation y_{ij} est donc égale à la longueur du vecteur $\boldsymbol{b}_j$, multipliée par la longueur de la projection du vecteur $\boldsymbol{a}_i$ sur le vecteur $\boldsymbol{b}_j$ ou encore à la longueur du vecteur $\boldsymbol{a}_i$, multipliée par la longueur de la projection du vecteur $\boldsymbol{b}_j$ sur le vecteur $\boldsymbol{a}_i$. Il en résulte que l'angle θ_{ij} entre les deux vecteurs $\boldsymbol{a}_i$ et $\boldsymbol{b}_j$ est donné par la relation :

$$\theta_{ij} = \arccos\left(\frac{< \boldsymbol{a}_i, \, \boldsymbol{b}_{j>}}{\|\boldsymbol{a_i}\|\|\boldsymbol{b_j}\|}\right) = \arccos\left(\frac{\sum_{s=1}^{2} a_{is} \, b_{js}}{\sqrt{\sum_{s=1}^{2} a_{is}^2 \; \sum_{s=1}^{2} b_{js}^2}}\right).$$

Deux vecteurs $\boldsymbol{a}_i$ et $\boldsymbol{b}_j$ forment par conséquent un angle droit lorsque $y_{ij} = 0$, puisque, dans ce cas, le produit scalaire est nul. Ils forment un angle aigu lorsque $y_{ij} > 0$ et ils forment un angle obtus si $y_{ij} < 0$.

3. L'angle formé par les deux vecteurs $\boldsymbol{a_i}$ et $\boldsymbol{a_{i'}}$ $(i' \neq i)$ ou $\boldsymbol{b_j}$ et $\boldsymbol{b_{j'}}$ $(j' \neq j)$ est égal à :

$$\theta_{ii'} = \arccos\left(\frac{\sum_{s=1}^{2} a_{is} \, a_{i's}}{\sqrt{\sum_{s=1}^{2} a_{is}^2 \; \sum_{s=1}^{2} a_{i's}^2}}\right) \text{ et } \theta_{jj'} = \arccos\left(\frac{\sum_{s=1}^{2} b_{js} \, b_{j's}}{\sqrt{\sum_{s=1}^{2} b_{js}^2 \; \sum_{s=1}^{2} b_{j's}^2}}\right).$$

2.3. Points-lignes et points-colonnes pour une matrice de rang quelconque

Si la matrice est de rang $r \quad (r > 2)$, les vecteurs $\boldsymbol{a}_i$ et $\boldsymbol{b}_j$ doivent être représentés dans un espace de dimension r et les propriétés vues ci-dessus restent valables lorsque les différentes sommes qui interviennent dans les relations sont étendues aux r termes.

Si une telle extension ne pose pas de problèmes théoriques, elle ne permet cependant plus une représentation graphique simple. Pour se limiter, par la suite, à une représentation graphique à deux dimensions, il peut se justifier de ne retenir que deux colonnes de $\boldsymbol{A}$ et de $\boldsymbol{B}$ et de négliger les $r-2$ autres colonnes. Les propriétés relatives aux vecteurs rappelées ci-dessus restent valables, sauf en ce qui concerne le produit scalaire des vecteurs $\boldsymbol{a}_i$ et $\boldsymbol{b}_j$. Celui-ci n'est, en effet, plus égal à y_{ij}, mais à une approximation de y_{ij}, $r-2$ termes ayant été négligés. Comme nous le verrons ci-dessous, dans la pratique on retiendra souvent les deux premières colonnes de $\boldsymbol{A}$ et de $\boldsymbol{B}$ et on négligera les $r-2$ dernières colonnes, mais la factorisation de $\boldsymbol{Y}$ sera optimisée, de manière à ce que l'approximation des y_{ij} soit aussi bonne que possible.

Pour ne pas alourdir inutilement les notations, nous ne ferons plus, par la suite la distinction entre le cas où $\boldsymbol{Y}$ est de rang 2 et le cas où $\boldsymbol{Y}$ est de

rang supérieur à 2. Nous donnerons les formules générales pour les matrices de rang r, étant entendu que, dans la pratique, on se limitera le plus souvent aux deux premières colonnes de $\boldsymbol{A}$ et de $\boldsymbol{B}$, ce qui revient à dire qu'on n'étudie pas la matrice $\boldsymbol{Y}$, mais une approximation de celle-ci, notée $\hat{\boldsymbol{Y}}_{(2)}$ ou encore, plus simplement, $\hat{\boldsymbol{Y}}$, qui est une matrice de rang 2.

Si, pour l'exemple examiné précédemment, on élimine la dernière colonne des matrices $\boldsymbol{A}$ et $\boldsymbol{B}$, le produit matriciel ne redonne pas la matrice $\boldsymbol{Y}$ mais une approximation de $\boldsymbol{Y}$:

$$\hat{\boldsymbol{Y}} = \begin{bmatrix} 3,11 & -1,31 \\ 1,56 & -1,37 \\ 2,31 & 0,10 \\ 4,46 & 1,34 \end{bmatrix} \begin{bmatrix} 4,47 & 2,50 & 3,33 \\ -0,76 & 2,12 & -0,57 \end{bmatrix} = \begin{bmatrix} 14,92 & 5,00 & 11,11 \\ 8,02 & 1,00 & 5,97 \\ 10,26 & 6,00 & 7,65 \\ 18,92 & 14,00 & 14,11 \end{bmatrix},$$

qui, compte tenu du choix de $\boldsymbol{A}$ et $\boldsymbol{B}$, est excellente, la somme des carrés des écarts entre les valeurs initiales et les valeurs approchées étant égale à :

$$\sum_{i=1}^{4}\sum_{j=1}^{3}(y_{ij} - \hat{y}_{ij})^2 = 0,23.$$

La figure 1 donne la représentation des lignes de $\boldsymbol{A}$ et de $\boldsymbol{B}$ après suppression de la dernière colonne de chacune de ces matrices. Les points représentant les lignes de $\boldsymbol{A}$ (points-lignes) sont désignés par les symboles L_1 à L_4 et les points représentant les lignes de $\boldsymbol{B}$ (points-colonnes) sont représentés par les symboles C_1 à C_3.

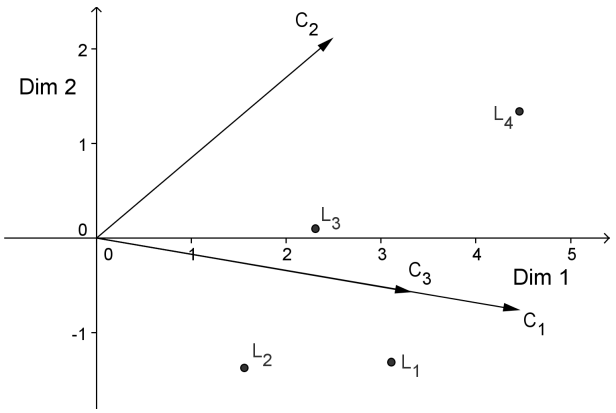

Figure 1 – Biplot de la matrice $\hat{\boldsymbol{Y}}$ (L_1 à L_4 : points-lignes ; C_1 à C_3 : points-colonnes).

Ce graphique montre immédiatement que les vecteurs C_1 et C_3 ont pratiquement la même direction : la première et la troisième ligne de B (limitées aux deux premières colonnes) sont donc à peu près proportionnelles. On constate aussi que le produit scalaire de L_4 et C_1 est le plus grand, ce qui signifie que

l'élément $\hat{y}_{41}$ de $\hat{\boldsymbol{Y}}$ est le plus grand. On note encore que tous les produits scalaires d'un point-ligne avec un point-colonne sont positifs, chaque angle étant inférieur à 90 ^{0} ; la matrice $\hat{\boldsymbol{Y}}$ ne contient donc que des éléments positifs. L'angle formé par les vecteurs L_2 et C_2 est proche de 90^0 ; la valeur $\hat{y}_{22}$ est donc très faible par rapport aux autres valeurs. Dans la mesure où $\hat{\boldsymbol{Y}}$ est proche de $\boldsymbol{Y}$, les commentaires faits à propos des valeurs $\hat{y}_{ij}$ restent valables pour les y_{ij}.

3. DÉCOMPOSITION PAR VALEURS SINGULIÈRES

3.1. Optimisation de la factorisation

Au paragraphe 2.1, nous avons vu qu'une matrice de rang 2 peut être représentée sous la forme d'un biplot, qui est une représentation simultanée d'un "effet ligne" et d'un "effet colonne". Ce biplot est réalisé à partir de la factorisation de $\boldsymbol{Y}$ sous la forme d'un produit de deux matrices. Lorsque la matrice est de rang $r \quad (r > 2)$, le biplot construit à partir des deux premières colonnes de $\boldsymbol{A}$ et de $\boldsymbol{B}$ est la représentation d'une matrice de rang 2, notée $\hat{\boldsymbol{Y}}_{(2)}$ ou $\hat{\boldsymbol{Y}}$, qui est une approximation de $\boldsymbol{Y}$. Puisque la factorisation de $\boldsymbol{Y}$ par le produit de deux matrices $\boldsymbol{A}$ et $\boldsymbol{B'}$ n'est pas unique, il y a évidemment intérêt à choisir, parmi l'ensemble des solutions possibles, celle qui, lorsqu'on ne retient que les deux premières colonnes de $\boldsymbol{A}$ et de $\boldsymbol{B}$, donne la meilleure approximation de $\boldsymbol{Y}$. La décomposition par valeurs singulières de la matrice $\boldsymbol{Y}$ donne la solution à ce problème.

La *décomposition par valeurs singulières*[1] d'une matrice $\boldsymbol{Y}$, de dimension $n \times p$ et de rang r consiste en la factorisation suivante [Eckart et Young, 1936] :

$$\boldsymbol{Y} = \boldsymbol{ULV'},$$

où $\boldsymbol{U}$ est une matrice de dimensions $n \times r$, $\boldsymbol{L}$ une matrice de dimensions $r \times r$ et $\boldsymbol{V}$ une matrice de dimensions $p \times r$. $\boldsymbol{U}$ et $\boldsymbol{V}$ sont des matrices orthonormées, c'est-à-dire telles que :

$$\boldsymbol{U'U} = \boldsymbol{I} \text{ et } \boldsymbol{V'V} = \boldsymbol{I}.$$

De plus, $\boldsymbol{U}$ est la matrice dont les colonnes sont les vecteurs propres, normés à l'unité, associés aux valeurs propres non nulles de $\boldsymbol{YY'}$; $\boldsymbol{V}$ est la matrice dont les colonnes sont les vecteurs propres normés à l'unité, associés aux valeurs propres non nulles de $\boldsymbol{Y'Y}$; $\boldsymbol{L}$ est la matrice diagonale dont les éléments diagonaux sont les valeurs singulières de $\boldsymbol{Y}$, c'est-à-dire les racines carrées des valeurs propres non nulles de $\boldsymbol{Y'Y}$ ou de $\boldsymbol{YY'}$, ces deux matrices ayant des valeurs propres identiques.

La décomposition par valeurs singulières factorise $\boldsymbol{Y}$ sous la forme d'un produit de trois matrices et nous verrons, au paragraphe suivant, comment celles-ci peuvent être utilisées pour une représentation par biplot.

1. En anglais : singular-value decomposition.

Pour la matrice $\boldsymbol{Y}$ présentée au paragraphe précédent, la décomposition conduit aux résultats suivants :

$$\boldsymbol{U} = \begin{bmatrix} 0,51 & -0,56 & 0,30 \\ 0,26 & -0,59 & -0,07 \\ 0,38 & 0,04 & -0,91 \\ 0,73 & 0,58 & 0,29 \end{bmatrix}, \quad \boldsymbol{V} = \begin{bmatrix} 0,73 & -0,33 & 0,60 \\ 0,41 & 0,91 & 0,00 \\ 0,55 & -0,24 & -0,80 \end{bmatrix}$$

$$\boldsymbol{L} = \begin{bmatrix} 37,36 & 0 & 0 \\ 0 & 5,40 & 0 \\ 0 & 0 & 0,48 \end{bmatrix}.$$

Les matrices $\boldsymbol{U}$ et $\boldsymbol{V}$ sont liées par les relations suivantes :

$$\boldsymbol{U} = \boldsymbol{YVL}^{-1} \qquad \text{et} \quad \boldsymbol{V} = \boldsymbol{Y'UL}^{-1},$$

ou encore, pour une colonne k particulière :

$$\boldsymbol{u}_k = \frac{1}{l_k}\boldsymbol{Y}\boldsymbol{v}_k \qquad \text{et} \quad \boldsymbol{v}_k = \frac{1}{l_k}\boldsymbol{Y'}\boldsymbol{u}_k \qquad (k = 1, ..., r).$$

Les matrices $\boldsymbol{U}$, $\boldsymbol{L}$ et $\boldsymbol{V}$ ont encore les propriétés suivantes :

$$(\boldsymbol{UL})(\boldsymbol{UL})' = \boldsymbol{ULL'U'} = \boldsymbol{UL}^2\boldsymbol{U}' = \boldsymbol{YY'}$$

$$(\boldsymbol{VL})(\boldsymbol{VL})' = \boldsymbol{VLL'V'} = \boldsymbol{VL}^2\boldsymbol{V}' = \boldsymbol{Y'Y},$$

$\boldsymbol{L}^2$ étant la matrice diagonale des valeurs propres non nulles de $\boldsymbol{Y'Y}$ et de $\boldsymbol{YY'}$.

3.2. Reconstitution de Y

La factorisation de $\boldsymbol{Y}$ peut s'écrire sous la forme d'une somme de produits de vecteurs :

$$\boldsymbol{Y} = \boldsymbol{ULV'} = \sum_{k=1}^{r} l_k \boldsymbol{u}_k \boldsymbol{v'}_k = \boldsymbol{Y}_1 + \boldsymbol{Y}_2 + ... + \boldsymbol{Y}_r.$$

Ainsi, pour l'exemple, on a :

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

$$+ \begin{bmatrix} 0,09 & 0,00 & -0,11 \\ -0,02 & 0,00 & 0,03 \\ -0,26 & 0,00 & 0,35 \\ 0,08 & 0,00 & -0,11 \end{bmatrix}.$$

Chaque matrice $\boldsymbol{Y}_k \quad (k = 1, ..., r)$ est une matrice de rang 1 et la somme de $q \quad (q \leq r)$ de ces matrices donne une matrice de rang q. L'importance de chacune de ces matrices dans la reconstitution de $\boldsymbol{Y}$ est décroissante : $\boldsymbol{Y}_1$ est la matrice de rang 1 qui donne la meilleure approximation de $\boldsymbol{Y}$; $\boldsymbol{Y}_1 + \boldsymbol{Y}_2$ est la matrice de rang 2 qui donne la meilleure approximation de $\boldsymbol{Y}$ et ainsi de suite, la qualité de l'approximation se mesurant par la somme des carrés des écarts entre les y_{ij} et les y_{ij} approchés, notés $\hat{y}_{ij}$, pour l'ensemble des couples ij.

La qualité relative de la reconstitution de $\boldsymbol{Y}$ à partir de $\boldsymbol{Y}_1 + \boldsymbol{Y}_2 + ... + \boldsymbol{Y}_q$ est mesurée par le rapport :

$$(l_1^2 + l_2^2 + ... + l_q^2)/(l_1^2 + l_2^2 + ... + l_r^2),$$

qui est égal à :

$$\sum_{i=1}^{n}\sum_{j=1}^{p} \hat{y}_{ij}^2 / \sum_{i=1}^{n}\sum_{j=1}^{p} y_{ij}^2.$$

Plus ce rapport est proche de l'unité, plus l'approximation de $\boldsymbol{Y}$ est bonne.

Ainsi, pour l'exemple ci-dessus, si on néglige la dernière colonne des matrices $\boldsymbol{U}$ et $\boldsymbol{V}$, on a :

$$\hat{\boldsymbol{Y}}_{(2)} = \boldsymbol{Y}_1 + \boldsymbol{Y}_2,$$

et l'approximation de $\boldsymbol{Y}$ par $\hat{\boldsymbol{Y}}_{(2)}$ est excellente car le rapport :

$$(37,36^2 + 5,40^2)/(37,36^2 + 5,40^2 + 0,48^2) = 0,9998$$

est pratiquement égal à l'unité.

4. COORDONNÉES DES POINTS DANS LE BIPLOT

4.1. Différentes solutions

La décomposition par valeurs singulières factorise la matrice $\boldsymbol{Y}$ en un produit de trois matrices.

D'autre part, nous avons vu, au paragraphe 2, que le biplot repose sur la factorisation de $\boldsymbol{Y}$ en deux matrices $\boldsymbol{A}$ et $\boldsymbol{B}$. Ces deux matrices vont être définies à partir de la décomposition par valeurs singulières.

On a :

$$\boldsymbol{Y} = \boldsymbol{ULV'} = \boldsymbol{AB'},$$

et la détermination de $\boldsymbol{A}$ et $\boldsymbol{B}$ à partir de $\boldsymbol{U}$, $\boldsymbol{L}$ et $\boldsymbol{V}$ peut se faire de plusieurs manières. En fonction d'un paramètre α, on peut écrire la décomposition sous la forme suivante :

$$\boldsymbol{Y} = (\boldsymbol{UL}^{\alpha})(\boldsymbol{L}^{1-\alpha}\boldsymbol{V})' = \boldsymbol{AB'} \qquad (0 < \alpha < 1),$$

$\boldsymbol{L}^{\alpha}$ étant la matrice diagonale dont les éléments sont les valeurs λ_k^{α}.

Dans la pratique, on considère le plus souvent l'une des trois solutions suivantes, qui correspondent respectivement à $\alpha = 1, \alpha = 0$ et $\alpha = 1/2$:

$$(\boldsymbol{UL})(\boldsymbol{V})' = \boldsymbol{A}_1\boldsymbol{B}_1' \quad \text{avec} \quad \boldsymbol{A}_1 = \boldsymbol{UL} \text{ et } \boldsymbol{B}_1 = \boldsymbol{V}$$

$$(\boldsymbol{U})(\boldsymbol{LV})' = \boldsymbol{A}_2\boldsymbol{B}_2' \quad \text{avec} \quad \boldsymbol{A}_2 = \boldsymbol{U} \text{ et } \boldsymbol{B}_2 = \boldsymbol{VL}$$

et

$$(\boldsymbol{UL}^{1/2})(\boldsymbol{L}^{1/2}\boldsymbol{V})' = \boldsymbol{A}_3\boldsymbol{B}_3' \quad \text{avec} \quad \boldsymbol{A}_3 = \boldsymbol{UL}^{1/2} \text{ et } \boldsymbol{B}_3 = \boldsymbol{VL}^{1/2}.$$

4.2. Respect des distances et des angles

Considérons d'abord le cas d'une matrice $\boldsymbol{Y}$ de rang 2 et examinons la factorisation pour $\alpha = 1$. D'après les propriétés données au paragraphe 3.1, on a :

$$\boldsymbol{A}_1\boldsymbol{A}_1' = \boldsymbol{UL}(\boldsymbol{UL})' = \boldsymbol{YY}'.$$

Les éléments diagonaux des matrices $\boldsymbol{A}_1\boldsymbol{A}_1'$ et $\boldsymbol{YY}'$ correspondent aux sommes des carrés des éléments des lignes des matrices $\boldsymbol{A}_1$ et $\boldsymbol{Y}$. On a donc :

$$\boldsymbol{y}_i\boldsymbol{y}_i' = \boldsymbol{a}_i\boldsymbol{a}_i' = \sum_{j=1}^{p} y_{ij}^2.$$

La longueur du vecteur correspondant à un point-ligne du biplot est donc égale à la longueur du vecteur correspondant à la même ligne de la matrice $\boldsymbol{Y}$, si on représente les lignes de $\boldsymbol{Y}$ dans l'espace des colonnes de $\boldsymbol{Y}$.

D'autre part, pour un élément hors de la diagonale de $\boldsymbol{A}_1\boldsymbol{A}_1'$ et de $\boldsymbol{YY}'$, on a aussi :

$$\boldsymbol{y}_i\boldsymbol{y}_{i'}' = \boldsymbol{a}_i\boldsymbol{a'}_{i'} \qquad (i \neq i').$$

Il en résulte que l'angle formé par deux points-lignes du biplot est égal à l'angle formé par les deux vecteurs $\boldsymbol{y}_i$ et $\boldsymbol{y}_{i'}$ si on représente ces deux vecteurs dans l'espace des p colonnes :

$$\theta_{ii'} = \arccos\left(\frac{<\boldsymbol{y}_i,\ \boldsymbol{y}_{i'}>}{\|\boldsymbol{y_i}\|\|\boldsymbol{y_{i'}}\|}\right) = \arccos\left(\frac{<\boldsymbol{a}_i,\ \boldsymbol{a}_{i'}>}{\|\boldsymbol{a_i}\|\|\boldsymbol{a_{i'}}\|}\right),$$

ou encore :

$$\theta_{ii'} = \arccos\left(\frac{\sum_{j=1}^{p} y_{ij}\ y_{i'j}}{\sqrt{\sum_{j=1}^{p} y_{ij}^2 \sum_{j=1}^{p} y_{i'j}^2}}\right) = \arccos\left(\frac{\sum_{s=1}^{2} a_{is}\ a_{i's}}{\sqrt{\sum_{s=1}^{2} a_{is}^2 \sum_{s=1}^{2} a_{i's}^2}}\right).$$

Le respect des longueurs et des angles entraîne le respect des distances euclidiennes entre les points-lignes du biplot par rapport aux distances euclidiennes entre les lignes de $\boldsymbol{Y}$.

Cette propriété du respect des distances euclidiennes et des angles n'est, par contre, pas vérifiée pour les points-colonnes, pour lesquels :

$$\boldsymbol{B}_1'\boldsymbol{B}_1 = \boldsymbol{I} \neq \boldsymbol{Y'Y}.$$

Par symétrie, la factorisation avec $\alpha = 0$ conduit au résultat suivant :

$$\boldsymbol{B}_2\boldsymbol{B}_2' = (\boldsymbol{VL})(\boldsymbol{VL})' = \boldsymbol{Y'Y}.$$

Pour cette factorisation, le biplot conserve les angles et les longueurs des vecteurs relatifs aux points-colonnes, mais pas des vecteurs relatifs aux points-lignes.

Enfin, la factorisation avec $\alpha = 1/2$ ne respecte les longueurs et les angles, ni pour les vecteurs relatifs aux points-lignes, ni pour les vecteurs relatifs aux points-colonnes.

En conclusion, la factorisation avec $\alpha = 1$ privilégie les lignes, la factorisation avec $\alpha = 0$ privilégie les colonnes et la factorisation avec $\alpha = 1/2$ donne un poids égal aux lignes et aux colonnes. Ces factorisations sont parfois appelées factorisation JK ou RMP (*row metric preserving*), GH ou CMP (*column metric preserving*) et SYM (*Symetric*).

Lorsque $\boldsymbol{Y}$ n'est pas une matrice de rang 2, nous avons vu que les biplots ne sont plus des représentations de $\boldsymbol{Y}$ mais des représentations d'une approximation de rang 2 de $\boldsymbol{Y}$. Dans ces conditions, les longueurs et les angles ne sont plus exactement conservés pour les points-lignes de $\boldsymbol{Y}$ lorsque $\alpha = 1$, ni pour les points-colonnes de $\boldsymbol{Y}$ lorsque $\alpha = 0$.

Pour l'exemple considéré précédemment, et en négligeant la troisième valeur propre, on obtient pour $\alpha = 1$:

$$\boldsymbol{A}_1 = \begin{bmatrix} 19,02 & -3,05 \\ 9,53 & -3,18 \\ 14,13 & 0,24 \\ 27,26 & 3,11 \end{bmatrix} \quad \text{et} \quad \boldsymbol{B}_1 = \begin{bmatrix} 0,73 & -0,33 \\ 0,41 & 0,91 \\ 0,55 & -0,24 \end{bmatrix}$$

et pour $\alpha = 0$:

$$\boldsymbol{A}_2 = \begin{bmatrix} 0,51 & -0,56 \\ 0,26 & -0,59 \\ 0,38 & 0,04 \\ 0,73 & 0,58 \end{bmatrix} \quad \text{et} \quad \boldsymbol{B}_2 = \begin{bmatrix} 27,33 & -1,78 \\ 15,29 & 4,93 \\ 20,37 & -1,31 \end{bmatrix}.$$

Pour $\alpha = 1/2$, on obtient la factorisation donnée au paragraphe 2.1.

Les représentations sous forme de biplots pour les factorisations avec $\alpha = 1$ et $\alpha = 0$ sont données aux figures 2 et 3 et, pour $\alpha = 1/2$, à la figure 1, examinée au paragraphe 2.3.

On constate que, dans la figure 2, qui privilégie la représentation des lignes, les points-colonnes ont des coordonnées très faibles, ce qui rend le graphique peu

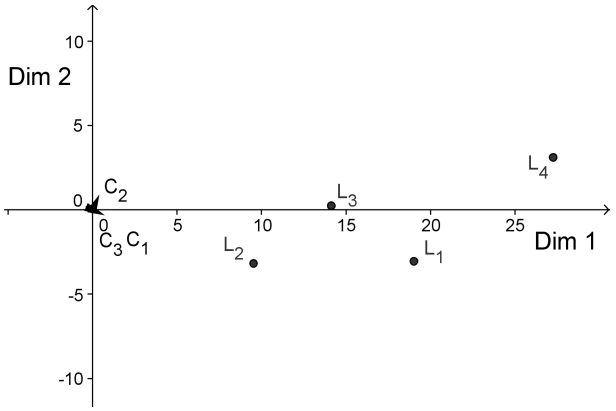

Figure 2 – Biplot de la matrice $\hat{\boldsymbol{Y}}$, factorisation avec $\alpha = 1$ (L_1 à L_4 : points-lignes ; C_1 à C_3 : points-colonnes).

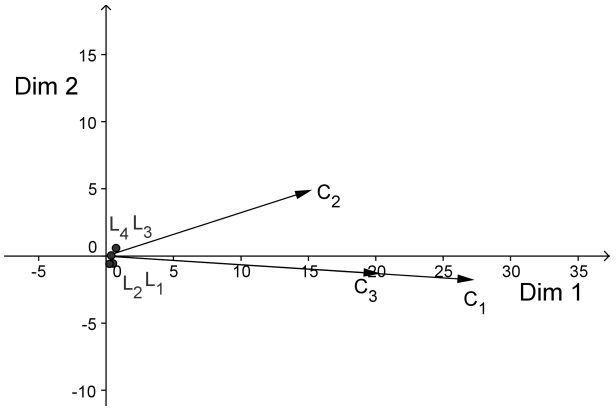

Figure 3 – Biplot de la matrice $\hat{\boldsymbol{Y}}$, factorisation avec $\alpha = 0$ (L_1 à L_4 : points-lignes ; C_1 à C_3 : points-colonnes) .

lisible pour les points-colonnes. On a, par contre, la situation inverse pour la figure 3, les points-lignes étant alors peu visibles. Enfin, dans la figure 1, qui consiste en un compromis, les deux types de points sont visibles simultanément. Cette discordance d'échelle pour les points-lignes et les points-colonnes lorsque $\alpha = 1$ ou $\alpha = 0$ sera systématique, dès que le valeurs singulières de $\boldsymbol{Y}$ seront différentes de l'unité.

L'examen de la figure 2, qui respecte les longueurs et les angles pour les points-lignes, montre par exemple que le vecteur pour L_4 est à peu près deux fois plus long que le vecteur pour L_3. Cela signifie donc que la moyenne quadratique des données de la ligne 4 de $\hat{\boldsymbol{Y}}$ est approximativement le double de la moyenne quadratique des données de la ligne 3 de $\hat{\boldsymbol{Y}}$.

Effectivement, les racines carrées des sommes des carrés des éléments des

lignes 3 et 4 de $\hat{\boldsymbol{Y}}$ sont respectivement égales à (paragraphe 2.3) :

$$\sqrt{10,26^2 + 6,00^2 + 7,65^2} = 14,13$$

et

$$\sqrt{18,92^2 + 14,00^2 + 14,11^2} = 27,44,$$

ces valeurs étant égales aux longueurs des vecteurs des points-lignes, pour les lignes 3 et 4 de la matrice $\boldsymbol{A}_1$:

$$\sqrt{14,13^2 + 0,24^2} = 14,13$$

et

$$\sqrt{27,26^2 + 3,11^2} = 27,44.$$

On voit immédiatement aussi que les moyennes quadratiques des observations des lignes s'ordonnent de la manière croissante suivante : 2, 3, 1 et 4.

La figure 3, qui respecte les longueurs et les distances pour les points-colonnes, montre par exemple que la colonne 2 présente des valeurs globalement plus faibles que la colonne 1 et que la colonne 3 présente des valeurs intermédiaires, les carrés des longueurs des vecteurs dans la figure 3 étant égaux aux sommes des carrés des éléments des colonnes de $\hat{\boldsymbol{Y}}$. Par exemple, pour la première colonne de $\hat{\boldsymbol{Y}}$ (paragraphe 2.3) et la première ligne de $\boldsymbol{B}_2$, on a :

$$14,92^2 + 8,02^2 + 10,26^2 + 18,92^2 = 27,33^2 + (-1,78^2) = 750,1.$$

Cette valeur est aussi très proche de la somme des carrés de la première colonne de $\boldsymbol{Y}$, puisque $\hat{\boldsymbol{Y}}$ est très proche de $\boldsymbol{Y}$ (paragraphe 3.2).

5. BIPLOT ET ANALYSE EN COMPOSANTES PRINCIPALES

5.1. Présentation classique de l'analyse en composantes principales

L'analyse en composantes principales repose sur le calcul des valeurs et des vecteurs propres de la matrice des variances et covariances $\boldsymbol{S}$ ou de la matrice de corrélation $\boldsymbol{R}$ des variables. Les deux méthodes se différencient par la standardisation préliminaire des variables de départ. Ainsi, pour l'analyse en composantes principales basée sur la matrice des variances et covariances, on considère que la matrice de départ, que nous notons $\boldsymbol{X}$, est la matrice des observations centrées obtenues en retranchant de chaque observation d'une colonne, c'est-à-dire d'une variable, la moyenne de la colonne. Pour une analyse sur la matrice de corrélation, on considère que la matrice $\boldsymbol{X}$ est constituée de variables centrées et réduites : on retranche des observations d'une colonne la moyenne de la colonne et on divise le résultat par l'écart-type de la colonne. Par la suite, nous considérons essentiellement le cas de la matrice de corrélation, qui correspond à la situation la plus fréquente en pratique.

Soit μ_k^2 et $\boldsymbol{c}_k$ les valeurs propres et les vecteurs propres de la matrice $\boldsymbol{R}$. Soit $\boldsymbol{C}$ la matrice obtenue par la juxtaposition des vecteurs propres. Les composantes principales sont obtenues par la relation :

$$\boldsymbol{Z} = \boldsymbol{XC},$$

et le graphique des individus dans les différents plans factoriels correspond au diagramme de dispersion de z_{ik} et $z_{ik'}$ $\quad (k \neq k')$.

Pour la représentation des variables dans les cercles de corrélation, on calcule la matrice de corrélation $\boldsymbol{W}$ entre les variables initiales et les composantes principales, en multipliant les vecteurs propres $\boldsymbol{c}_k$ par la racine carrée de la valeur propre correspondante μ_k, soit :

$$\boldsymbol{W} = \boldsymbol{CM},$$

$\boldsymbol{M}$ étant la matrice diagonale dont les éléments diagonaux sont les μ_k. Dans le plan factoriel (k, k'), la variable j a comme coordonnées les valeurs w_{jk} et $w_{jk'}$.

A titre d'illustration, nous reprenons l'exemple proposé par Hartigan (1975), concernant les teneurs en protéines, graisse et lactose du lait de 16 mammifères. L'analyse en composantes principales de ces données a été détaillée antérieurement (Palm, 1998). Le lecteur y trouvera les données et l'ensemble des résultats de l'analyse, ainsi que des éléments d'interprétation. Nous n'envisageons ici que les aspects plus directement en relation avec les biplots.

Le tableau 1 reprend les données centrées et réduites notées x_{i1}, x_{i2} et x_{i3}. La juxtaposition de ces trois colonnes donne la matrice $\boldsymbol{X}$. Le tableau reprend également les valeurs des composantes principales, notées z_{i1}, z_{i2} et z_{i3}.

La matrice de corrélation des trois variables est égale à :

$$\boldsymbol{R} = \begin{bmatrix} 1,000 & 0,897 & -0,938 \\ 0,897 & 1,000 & -0,865 \\ -0,938 & -0,865 & 1,000 \end{bmatrix}.$$

Les valeurs propres de cette matrice valent :

$$\mu_1^2 = 2,800 \qquad \mu_2^2 = 0,142 \quad \text{et} \quad \mu_3^2 = 0,058,$$

et la matrice des vecteurs propres s'écrit :

$$\boldsymbol{C} = \begin{bmatrix} -0,585 & -0,233 & 0,777 \\ -0,569 & 0,801 & -0,188 \\ 0,578 & 0,552 & 0,601 \end{bmatrix}.$$

Les valeurs des composantes principales du tableau 1 s'obtiennent par le produit $\boldsymbol{XC}$ et le graphique des individus dans le premier plan factoriel est simplement le diagramme de dispersion de z_{i1} et z_{i2}.

Tableau 1 – Teneurs en protéines, graisse et lactose du lait de 16 mammifères : données centrées et réduites et valeurs des composantes principales.

Code	Nom	x_{i1}	x_{i2}	x_{i3}	z_{i1}	z_{i2}	z_{i3}
a	ânesse	$-1,354$	$-1,024$	$1,263$	$2,105$	$0,193$	$-0,100$
b	baleine	$1,332$	$1,858$	$-1,615$	$-2,770$	$0,285$	$-0,285$
c	biche	$1,132$	$1,639$	$-0,990$	$-2,167$	$0,502$	$-0,023$
d	brebis	$-0,239$	$-0,297$	$0,325$	$0,496$	$-0,002$	$0,065$
e	buffle	$-0,154$	$-0,078$	$0,325$	$0,322$	$0,152$	$0,090$
f	chamelle	$-0,839$	$-0,733$	$0,387$	$1,132$	$-0,177$	$-0,282$
g	cobaye	$0,275$	$-0,180$	$-0,927$	$-0,594$	$-0,720$	$-0,310$
h	jument	$-1,096$	$-1,083$	$1,701$	$2,241$	$0,328$	$0,374$
i	lama	$-0,725$	$-0,762$	$0,888$	$1,371$	$0,049$	$0,113$
j	lapine	$1,675$	$0,679$	$-1,428$	$-2,191$	$-0,636$	$0,316$
k	mule	$-1,268$	$-0,966$	$0,825$	$1,768$	$-0,022$	$-0,308$
l	rate	$0,789$	$0,606$	$-0,551$	$-1,125$	$-0,004$	$0,168$
m	renarde	$0,046$	$-0,369$	$0,450$	$0,443$	$-0,058$	$0,376$
n	renne	$1,218$	$1,727$	$-1,052$	$-2,303$	$0,517$	$-0,010$
o	truie	$0,189$	$-0,486$	$-0,301$	$-0,008$	$-0,599$	$0,057$
p	zèbre	$-0,982$	$-0,529$	$0,700$	$1,280$	$0,192$	$-0,243$

En multipliant $\boldsymbol{C}$ par la matrice diagonale dont les éléments sont les μ_k, on obtient les corrélations des variables initiales et des composantes principales :

$$\boldsymbol{W} = \begin{bmatrix} -0,585 & -0,233 & 0,777 \\ -0,569 & 0,801 & -0,188 \\ 0,578 & 0,552 & 0,601 \end{bmatrix} \begin{bmatrix} \sqrt{2,800} & 0 & 0 \\ 0 & \sqrt{0,142} & 0 \\ 0 & 0 & \sqrt{0,058} \end{bmatrix}$$

$$= \begin{bmatrix} -0,979 & -0,088 & 0,186 \\ -0,952 & 0,301 & -0,045 \\ 0,968 & 0,208 & 0,144 \end{bmatrix}.$$

Dans le premier plan factoriel, les variables protéines, graisse et lactose auront donc comme coordonnées les valeurs suivantes :

$$(-0,979,\ -0,088),\ \ (-0,952,\ 0,301)\ \text{ et }\ (0,968,\ 0,208).$$

5.2. Relations avec les représentations par biplots

Selon que $\boldsymbol{X}$ est la matrice des variables centrées ou des variables centrées réduites, la quantité :

$$(\boldsymbol{X'X})/(n-1)$$

représente la matrice de variances et covariances ou la matrice de corrélation. C'est donc la matrice dont on calcule les valeurs propres et les vecteurs propres.

Si on divise par $\sqrt{n-1}$ tous les éléments de $\boldsymbol{X}$:

$$\boldsymbol{Y} = \boldsymbol{X}/\sqrt{n-1},$$

alors la matrice $\boldsymbol{Y'Y}$ est directement la matrice de variances et covariances ou la matrice de corrélation, selon la définition de $\boldsymbol{X}$.

Dès lors, si on effectue la décomposition par valeurs singulières de $\boldsymbol{Y}$:

$$\boldsymbol{Y} = \boldsymbol{ULV'},$$

la matrice $\boldsymbol{L}$ est identique à la matrice $\boldsymbol{M}$ et la matrice $\boldsymbol{V}$ est identique à la matrice $\boldsymbol{C}$.

Il en résulte que le produit $\boldsymbol{YV}$ est égal, à une constante près, à la matrice $\boldsymbol{Z}$:

$$\boldsymbol{YV} = \boldsymbol{Z}/\sqrt{n-1}.$$

En remplaçant $\boldsymbol{Y}$ par sa décomposition par valeurs singulières et en tenant compte du fait que $\boldsymbol{V}$ est une matrice orthonormée, on a :

$$\boldsymbol{Z}/\sqrt{n-1} = \boldsymbol{ULV'V} = \boldsymbol{UL}.$$

Les coordonnées des individus dans les graphiques établis en analyse en composantes principales sont donc identiques, à une constante près, égale pour toutes les dimensions, aux coordonnées du biplot lors de la factorisation de $\boldsymbol{Y}$ avec $\alpha = 1$ (paragraphe 4.1).

Du fait de l'égalité des matrices $\boldsymbol{C}$ et $\boldsymbol{V}$ d'une part, et $\boldsymbol{M}$ et $\boldsymbol{L}$ d'autre part, la matrice $\boldsymbol{W}$ qui donne les coordonnées des variables est égale à :

$$\boldsymbol{W} = \boldsymbol{VL} = \boldsymbol{CM}.$$

Les coordonnées des variables dans l'analyse en composantes principales sont donc égales aux coordonnées des points-colonnes du biplot obtenu par factorisation de $\boldsymbol{Y}$ avec $\alpha = 0$ (paragraphe 4.1).

Pour les données relatives aux mammifères, la matrice $\boldsymbol{Y}$ s'écrit :

$$\boldsymbol{Y} = \begin{bmatrix} -0,350 & -0,264 & 0,326 \\ 0,344 & 0,480 & -0,417 \\ \ldots & \ldots & \ldots \\ -0,254 & -0,137 & 0,181 \end{bmatrix}.$$

Elle comporte 16 lignes, mais seules les deux premières et la dernière lignes ont été reprises.

La décomposition par valeurs singulières de cette matrice donne les résultats suivants :

$$\boldsymbol{U} = \begin{bmatrix} 0,325 & -0,133 & 0,108 \\ -0,427 & -0,195 & 0,307 \\ \ldots & \ldots & \ldots \\ 0,198 & -0,132 & 0,262 \end{bmatrix},$$

$$\boldsymbol{L} = \begin{bmatrix} 1,674 & 0 & 0 \\ 0 & 0,376 & 0 \\ 0 & 0 & 0,239 \end{bmatrix}$$

et

$$\boldsymbol{V} = \begin{bmatrix} -0,585 & 0,233 & -0,777 \\ -0,569 & -0,801 & 0,188 \\ 0,578 & -0,552 & -0,601 \end{bmatrix}.$$

On constate tout d'abord que la matrice $\boldsymbol{V}$ est bien égale à la matrice des vecteurs propres de $\boldsymbol{R}$, notée $\boldsymbol{C}$ au paragraphe 5.1, à condition toutefois de changer le signe des deux dernières colonnes. Cette inversion de signe résulte du fait qu'en analyse en composantes principales le signe des vecteurs propres est arbitraire et on peut toujours multiplier par -1 tous les éléments d'un vecteur propre. Si les signes n'ont pas été changés au paragraphe 5.1, c'est pour respecter les résultats donnés dans l'étude antérieure [Palm, 1998]. On constate aussi que si on élève au carré les éléments diagonaux de $\boldsymbol{L}$ on retrouve les valeurs propres μ_k^2 de la matrice $\boldsymbol{R}$ (paragraphe 5.1).

D'autre part, si on multiplie la première colonne de $\boldsymbol{U}$ par 1,674, la deuxième colonne de $\boldsymbol{U}$ par 0,376 et la troisième colonne de $\boldsymbol{U}$ par 0,239, on retrouve, au facteur $\sqrt{15}$ près, les valeurs des composantes principales. Ainsi, pour le premier animal, on a :

$$\begin{aligned} (0,325)(1,674)(\sqrt{15}) &= 2,107 \\ (-0,133)(0,376)(\sqrt{15}) &= -0,194 \\ (0,108)(0,239)(\sqrt{15}) &= 0,100. \end{aligned}$$

Ces valeurs sont, aux erreurs d'arrondis près, égales aux valeurs des composantes principales données dans le tableau 1, si on modifie le signe des deux dernières composantes.

Enfin, si on multiplie la matrice $\boldsymbol{V}$ par $\boldsymbol{L}$, c'est-à-dire si on multiplie la première colonne de $\boldsymbol{V}$ par 1,674, la deuxième colonne de $\boldsymbol{V}$ par 0,376 et la troisième colonne de $\boldsymbol{V}$ par 0,239, on retrouve la matrice de corrélation $\boldsymbol{W}$ du paragraphe 5.1, si on inverse les signes des deux dernières colonnes.

5.3. Distances euclidiennes et distances de Mahalanobis

Nous avons vu, au paragraphe 4.2, que la factorisation de $\boldsymbol{Y}$ après décomposition par valeurs singulières avec $\alpha = 1$ conduit à une représentation graphique qui respecte les longueurs pour les points-lignes et les distances euclidiennes entre les points-lignes. Ainsi, pour le premier individu, la somme des carrés des éléments de la première ligne de $\boldsymbol{Y}$ est égale à :

$$(-0,350)^2 + (-0,264)^2 + 0,326^2 = 0,298,$$

et la somme des différences entre les deux premières lignes vaut :

$$(-0,350 - 0,344)^2 + (-0,264 - 0,480)^2 + (0,326 + 0,417)^2 = 1,587.$$

La première valeur est égale à la somme des carrés des éléments de la première ligne de $\boldsymbol{UL}$, soit :

$$[(0,325)(1,674)]^2 + [(-0,133)(0,376)]^2 + [(0,108)(0,239)]^2 = 0,299.$$

La deuxième valeur est égale à la somme des carrés des différences entre les deux premières lignes de $\boldsymbol{UL}$:

$$[(0,325+0,427)(1,674)]^2 + [(-0,133+0,195)(0,376)]^2$$

$$+[(0,108-0,307)(0,239)]^2 = 1,588.$$

Ces résultats montrent bien que, lorsque toutes les composantes principales sont retenues, les longueurs et les distances euclidiennes entre individus sont préservées. Bien entendu, si on néglige une ou plusieurs composantes, ce sont les distances pour le tableau approché $\hat{\boldsymbol{Y}}$ qui sont conservées.

Si, pour la représentation des individus, on avait retenu la factorisation avec $\alpha = 0$, les longueurs et les distances euclidiennes entre individus n'auraient pas été respectées. Par contre, on peut montrer que les carrés des longueurs des vecteurs relatifs aux points-lignes sont, au facteur $n-1$ près, les carrés des distances au sens de Mahalanobis des individus au centre de gravité. On peut démontrer en effet que, pour un individu i donné :

$$\boldsymbol{u_i u_i'} = \frac{1}{n-1} \boldsymbol{y_i S^{-1} y_i'},$$

$\boldsymbol{S}$ étant la matrice de variances et covariances des données. Contrairement aux distances euclidiennes, les distances de Mahalanobis tiennent compte des variances et covariances entre variables. Elles sont largement utilisées en analyse multivariée, en relation avec l'hypothèse de multinormalité des populations.

De même, les distances euclidiennes entre deux points-lignes $\boldsymbol{u_i}$ et $\boldsymbol{u_{i'}}$ dans le biplot établi pour $\alpha = 0$ sont, à une constante près, égales au carré des distances au sens de Mahalanobis entre les individus $\boldsymbol{y_i}$ et $\boldsymbol{y_{i'}}$:

$$(\boldsymbol{u_i} - \boldsymbol{u_{i'}})(\boldsymbol{u_i} - \boldsymbol{u_{i'}})' = \frac{1}{n-1} (\boldsymbol{y_i} - \boldsymbol{y_{i'}}) \boldsymbol{S^{-1}} (\boldsymbol{y_i} - \boldsymbol{y_{i'}})'.$$

Pour les données relatives aux mammifères, le carré de la longueur du vecteur-ligne pour le premier individu est égal à :

$$0,325^2 + (-0,133)^2 + (0,108)^2 = 0,135,$$

et le carré de la distance euclidienne entre les deux premiers individus vaut :

$$(0,325+0,427)^2 + (-0,133+0,195)^2 + (0,108-0,307)^2 = 0,609.$$

Si on calcule le carré de la distance de Mahalanobis pour le premier individu et le carré de la distance de Mahalanobis entre les deux premiers

individus, on trouve respectivement 2,020 et 9,137. En divisant par 15 ces deux valeurs, on retrouve bien 0,135 et 0,609.

Considérons maintenant la représentation des variables réalisée lors de l'analyse en composantes principales. Nous avons vu que cette représentation correspond à la représentation des points-colonnes du biplot pour $\alpha = 0$ (paragraphe 5.2). Dans l'espace factoriel complet, le carré de la distance d'un point-colonne j à l'origine des axes correspond à l'élément jj de la matrice (paragraphe 4.2) :

$$\boldsymbol{WW'} = \boldsymbol{VL}(\boldsymbol{VL})' = \boldsymbol{Y'Y}.$$

Si l'analyse en composantes principales est réalisée sur la matrice de corrélation, alors (paragraphe 5.2) :

$$\boldsymbol{Y'Y} = \boldsymbol{R},$$

et les éléments diagonaux sont égaux à l'unité. Dans l'espace factoriel complet, les points représentant les variables se trouvent donc sur une hypersphère de rayon unitaire. Dans un sous-espace factoriel de dimension 2, ils se trouvent sur ou à l'intérieur du cercle de rayon unitaire, appelé cercle de corrélation.

Si on examine l'angle formé par deux vecteurs-colonnes j et j', on a (paragraphe 2) :

$$\cos\theta_{jj'} = \frac{\sum_{s=1}^{r} w_{js}\, w_{j's}}{\sqrt{\sum_{s=1}^{r} w_{js}^2 \sum_{s=1}^{r} w_{j's}^2}} = \sum_{s=1}^{r} w_{js}\, w_{j's}.$$

puisque, comme on vient de le signaler, la longueur des vecteurs relatifs aux points-colonnes est égale à l'unité. Il en résulte que $\cos\theta_{jj'}$ est l'élément jj' de la matrice $\boldsymbol{WW'}$ et donc de $\boldsymbol{R}$, c'est-à-dire aussi la corrélation des variables j et j'. Deux points sont confondus si les variables correspondantes ont une corrélation égale à l'unité : ils forment un angle droit si la corrélation est nulle et ils forment un angle de 180 degrés si la corrélation est égale à -1. L'égalité entre le cosinus de l'angle et la corrélation des variables n'est cependant vérifiée que pour l'espace factoriel complet.

Les propriétés énoncées ci-dessus peuvent être appliquées aux données relatives aux mammifères. Pour les deux premières variables, c'est-à-dire pour les deux premières lignes de $\boldsymbol{VL}$ prises comme exemple, on a bien :

$$(-0,979)^2 + 0,088^2 + (-0,186)^2 \simeq 1,$$

$$(-0,952)^2 + (-0,301)^2 + 0,045^2 \simeq 1,$$

et

$$(-0,979)(-0,952) + (0,088)(-0,301) + (-0,186)(0,045) = 0,897,$$

cette dernière valeur étant la corrélation entre les deux premières variables, c'est-à-dire entre les teneurs en protéines et en graisse.

5.4. Représentation simultanée des variables et des individus.

En analyse en composantes principales, les individus et les variables font le plus souvent l'objet de représentations graphiques séparées. Et nous avons vu, au paragraphe 5.2, que le graphique des individus est équivalent à la représentation des lignes dans le biplot avec $\alpha = 1$ et que le cercle de corrélation correspond à la représentation des colonnes dans le biplot avec $\alpha = 0$.

Une représentation simultanée des individus et des variables par le biplot avec $\alpha = 1$ peut cependant présenter un intérêt pratique, si on calibre les axes correspondant aux variables.

La figure 4 donne le biplot pour l'exemple des mammifères. Les individus sont identifiés par leur code (tableau 1) et les variables par les symboles P pour protéines, G pour graisse et L pour lactose.

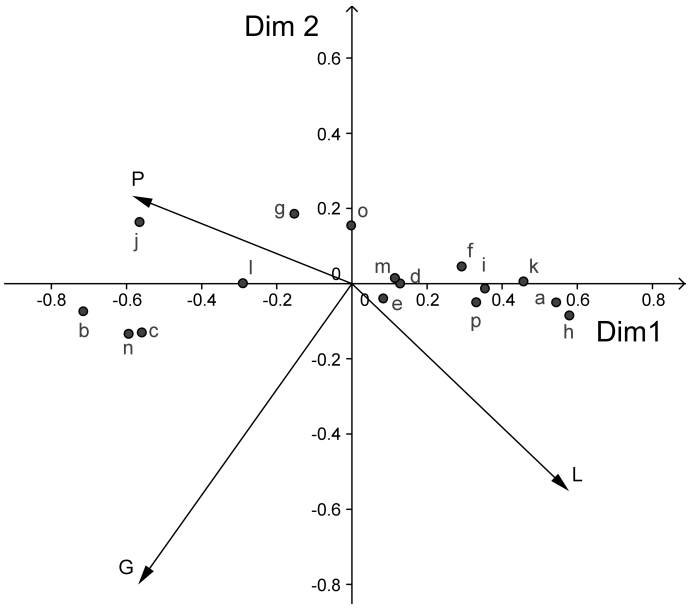

Figure 4 – Biplot de la matrice des données relatives aux mammifères, factorisation de $\boldsymbol{Y}$ avec $\alpha = 1$ ($a, b, ..., p$: code des animaux ; P : protéines, G : graisse et L : lactose).

A la figure 5, les vecteurs représentant les variables ont été remplacés par des axes gradués et, pour ne pas alourdir le graphique, seule l'ânesse, prise à titre d'exemple, a été identifiée par le symbole a. Si on projette un individu perpendiculairement sur un de ces axes , on obtient, comme coordonnée sur cet axe, la valeur approchée de la variable pour cet individu, que nous avons notée précédemment $\hat{y}_{ij}$. Ainsi par exemple, pour l'ânesse, on a respectivement pour l'axe "Protéines", "Graisse" et "Lactose", les valeurs suivantes :

$$\hat{y}_{11} = -0,329, \quad \hat{y}_{12} = -0,269 \quad \text{et} \quad \hat{y}_{13} = 0,342.$$

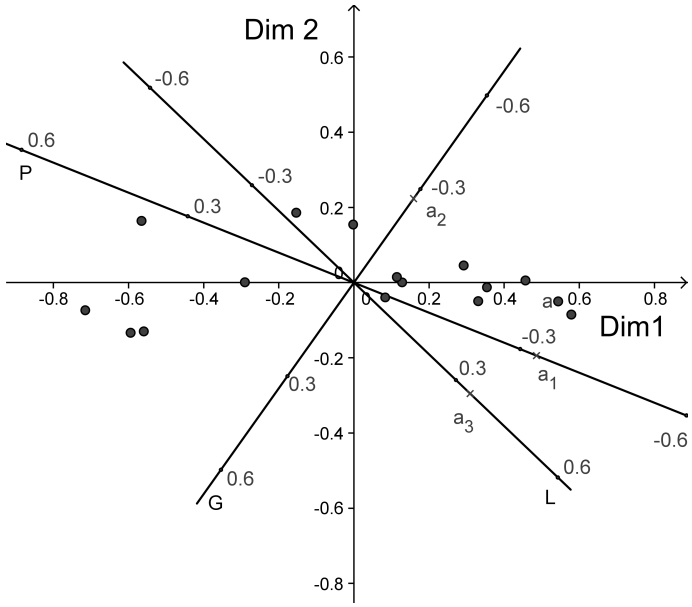

Figure 5 – Biplot de la matrice des données relatives aux mammifères, avec graduation des axes relatifs aux variables (a : ânesse ; a_1 : projection de a sur P ; a_2 : projection de a sur G ; a_3 : projection de a sur L).

Ces projections ont été matérialisées par les symboles a_1, a_2 et a_3. Les valeurs des projections sont très proches des valeurs de la première ligne de la matrice $\boldsymbol{Y}$ donnée au paragraphe 5.2 :

$$y_{11} = -0,350, \quad y_{12} = -0,264 \quad \text{et} \quad y_{13} = 0,326,$$

car, pour cet exemple, les deux premières composantes expliquent 98 % de la variabilité et la matrice $\boldsymbol{Y}$ est proche de $\hat{\boldsymbol{Y}}_{(2)}$.

La figure 5 peut donc être interprétée comme une extension multivariée d'un diagramme de dispersion : au lieu de deux axes perpendiculaires, on a plusieurs axes, nécessairement non perpendiculaires. Un tel graphique n'est cependant utile que si la matrice $\hat{\boldsymbol{Y}}_{(2)}$ est suffisamment proche de $\boldsymbol{Y}$, comme c'est le cas pour cet exemple.

Des informations concernant la détermination des graduations sur les axes sont données par Gower et *al.* (2011).

6. INFORMATIONS COMPLÉMENTAIRES

Dans cette note, nous avons présenté un outil permettant la visualisation graphique d'une matrice, généralement dans un espace de dimension réduite. La particularité de l'outil est de donner une représentation simultanée des lignes et des colonnes de la matrice, ce qui justifie le nom de biplot donné au graphique. La méthode repose sur la factorisation de la matrice sous la forme de deux matrices.

Cette factorisation est optimisée, de manière à ce que le sous-espace de dimension réduite donne la meilleure approximation de la matrice. L'objectif est atteint par l'utilisation de la décomposition par valeurs singulières :

$$\boldsymbol{Y} = \boldsymbol{ULV'}.$$

Les lignes et les colonnes de la matrice initiale $\boldsymbol{Y}$ sont alors représentées par des points dont les coordonnées sont les éléments des lignes et des matrices :

$$\boldsymbol{UL}^{\alpha} \text{ et } \boldsymbol{VL}^{\beta}.$$

Pour que le produit scalaire d'un vecteur correspondant à un point-ligne i et à un point-colonne j soit égal à l'élément ij de $\boldsymbol{Y}$, il faut que la somme des exposants de $\boldsymbol{L}$ dans les deux matrices ci-dessus soit égal à l'unité, c'est-à-dire que β soit égal à $1 - \alpha$.

Dans la pratique cependant, des représentations graphiques à partir de matrices pour lesquelles $\alpha + \beta \neq 1$ sont aussi utilisées, comme nous l'avons vu pour l'analyse en composantes principales, où $\alpha = \beta = 1$.

Une généralisation du biplot aux situations pour lesquelles $\alpha + \beta \neq 1$ est étudiée par Gower [2004] et la qualité de la représentation du biplot pour différents choix de α et β a été examinée par Gabriel [2002].

Dans cette note, nous nous sommes volontairement limités à une introduction aux biplots et à leur relation avec l'analyse en composantes principales. Il existe cependant aussi des relations entre les représentations par biplots et d'autres méthodes statistiques multivariées. Le lecteur trouvera des informations complémentaires dans les livres de Gower et Hand [1996] et de Gower *et al.* [2011].

BIBLIOGRAPHIE

Eckart C., Young G. [1936]. The approximation of a matrix by another of lower rank. *Psychometrika* 1, 211-218.

Gabriel K. R. [1971]. The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58(3), 453-467.

Gabriel K. R. [2002]. Goodness of fit of biplots and correspondance analysis. *Biometrika* 89(2), 423-436.

Gower J. C., Hand D. J. [1996]. *Biplots.* New York, Chapman and Hall, 277 p.

Gower J. C., Lubbe S., Roux N. [2011]. *Understanding biplots.* New York, Wiley, 463 p.

Hartigan J. A. [1975]. *Clustering algorithms.* New York, Wiley, 351 p.

Jackson J. E. [1991]. *A user's guide to principal components.* New York, Wiley, 569 p.

Khattree R., Naik D. N. [1995]. *Applied multivariate statistics with SAS software.* Cary N. C., SAS Institute Inc., 396 p.

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

Rencher A. C. [2002]. *Methods of multivariate analysis.* New York, Wiley, 708 p.

Seber G. A. F. [1984]. *Multivariate observations.* New York, Wiley, 686 p.

La collection

NOTES DE STATISTIQUE ET D'INFORMATIQUE

réunit divers travaux (documents didactiques, notes techniques, rapports de recherche, publications, etc.) émanant de l'Unité de Statistique, Informatique et Mathématique appliquées à la bioingénierie de l'Université de Liège – Gembloux Agro-Bio Tech et de l'Unité Systèmes agraires, Territoire et Technologies de l'Information du Centre wallon de Recherches agronomiques (Gembloux - Belgique).

La liste des notes disponibles peut être obtenue sur simple demande à l'adresse ci-dessous :

Université de Liège – Gembloux Agro-Bio Tech
Unité de Statistique, Informatique et Mathématique appliquées à la bioingénierie
Avenue de la Faculté d'Agronomie, 8
B-5030 GEMBLOUX (Belgique)
E-mail : sima.gembloux@ulg.ac.be

Plusieurs notes sont directement accessibles à l'adresse Web suivante, section Publications :

http://www.gembloux.ulg.ac.be/si/

En relation avec certaines notes, des programmes spécifiques sont également disponibles à la même adresse, section Macros.

Quelques titres récents sont cités ci-après :

PALM R., BROSTAUX Y. [2009]. Etude des séries chronologiques par les méthodes de décomposition. *Notes Stat. Inform.* (Gembloux) 2009/1, 17 p.

CHARLES C. [2011]. Introduction aux ondelettes. *Notes Stat. Inform*. (Gembloux) 2011/1, 22 p.

CHARLES C. [2011]. Introduction aux applications des ondelettes. *Notes Stat. Inform.* (Gembloux) 2011/2, 35 p.

PALM R., BROSTAUX Y. et CLAUSTRIAUX J. J. [2011]. Macros Minitab pour le choix d'une transformation pour la normalisation de variables. *Notes Stat. Inform*. (Gembloux) 2011/3, 15 p.

PALM R., BROSTAUX Y. [2011]. La régression logistique avec Minitab. *Notes Stat. Inform*. (Gembloux) 2011/4, 15 p.

PALM R., BROSTAUX Y., CLAUSTRIAUX J. J. [2011]. Inférence statistique et critères de qualité de l'ajustement en régression logistique binaire. *Notes Stat. Inform.* (Gembloux) 2011/5, 32 p.

CLAUSTRIAUX J. J., PALM R., FERRANDIS-VALLTERRA S., BROSTAUX Y. et PLANCHON V. [2012]. Tables de contingence à trois dimensions : aspects théoriques, applications et analogie avec l'analyse de la variance à trois critères de classification. *Notes Stat. Inform*. (Gembloux) 2012/1, 19 p.

D/2012/2371/2