JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

K. RUBEN GABRIEL

Le biplot - outil d'exploration de données multidimensionnelles

Journal de la société française de statistique, tome 143, nº 3-4 (2002), p. 5-55

<http://www.numdam.org/item?id=JSFS_2002__143_3-4_5_0>

© Société française de statistique, 2002, tous droits réservés.

L'accès aux archives de la revue « Journal de la société française de statistique » (http://publications-sfds.math.cnrs.fr/index.php/J-SFdS) implique l'accord avec les conditions générales d'utilisation (http://www.numdam.org/conditions). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
http://www.numdam.org/

LE BIPLOT – OUTIL D'EXPLORATION DE DONNÉES MULTIDIMENSIONNELLES

K. Ruben GABRIEL *

1. Introduction

Le biplot est un outil graphique pour visualiser des données arrangées en forme de matrice (Gabriel, 1971, 1982; Gower et Hand, 1996). Les graphiques ont des usages divers en statistique (voir, par exemple, Valois, 2000). L'un d'eux est l'exposition frappante de certains phénomènes connus que l'on souhaite visualiser au mieux. C'est un usage important pour la publicité, la gestion et l'instruction, mais on ne le discute pas ici. L'utilisation qu'on considère est au contraire celle du graphique « comme instrument d'une réflexion » (de Falguerolles, 2000), c'est-à-dire pour faciliter la découverte de phénomènes qu'on ne soupçonne pas avant. On présente des exemples montrant pourquoi et comment le biplot est bien utile pour cet usage de « reconnaissance, exploration and model building » (Friendly et Denis, 2000). C'est parce qu'il exhibe les données d'une manière simple et assez intuitive permettant à l'œil, qui est un instrument formidable de recherche quand il est lié à un cerveau actif, d'apercevoir des phénomènes inattendus, ce qui est difficile à faire avec des analyses formelles.

La description et la construction des biplots sont discutées dans la Section 2. La Section 3 explique l'usage de diverses métriques pour mettre en évidence des aspects d'intérêt comme les outliers, la dispersion entre échantillons, etc. Puis on montre comment utiliser le biplot pour diagnostiquer des structures qui ajustent bien les données (Section 4). Dans la Section 5 on discute les différents modes de représentation par biplot et ses relations avec le graphique introduit par Benzécri; on montre que la plupart de ces représentations se ressemblent si fortement que le choix entre elles est de peu d'importance. Pour expliquer le biplot on se fixe ici sur la représentation plane (bidimensionnelle) bien que des représentations en plus de dimensions soient possibles. Dans la Section 6 on revient à la représentation par biplot de données de types divers et au choix du bon nombre de dimensions pour un type ou l'autre : quand deux dimensions sont-elles suffisantes, quand faudrait-il faire une représentation multidimensionnelle? La Section 7 signale quelques problèmes qui restent à résoudre. Finalement (Section 8) on propose une définition plus générale du biplot et on présente quelques conclusions.

Les mathématiques utilisées pour la construction et l'utilisation des biplots sont celles de l'approximation des matrices en rang réduit. Notre approche est

* Department of Statistics, University of Rochester, Rochester, NY 14627, USA

algorithmique et inclut des formules analogues à celles de l'analyse statistique multidimensionnelle (Section 2.2) ce qui permet l'approximation graphique de certains tests statistiques (Section 3.7). Nous effleurons les problèmes de signification quand les questions de variabilité aléatoire se posent, mais notre abord est principalement la visualisation approximative des données dans un cadre exploratoire, en espérant que cela mène à des interprétations intelligentes.

2. Les biplots : exemple et construction

2.1. L'exemple des chiens et des loups

On commence par l'exemple d'un biplot (Figure 1) des données de $m = 6$ mesures sur $n = 43$ crânes, dont 30 de chiens, 12 de loups et un d'un animal non classifié *a priori* (Jambu, 1977). Les données centrées par colonnes sont rangées dans une matrice $\mathbf{Y}\boxed{(43 \times 6)}$. Chaque crâne $i = 1, \ldots, n$ est représenté sur le biplot par un point $\mathbf{a}_i$, étiqueté par la race de l'animal, et chaque mesure $j = 1, \ldots, m$ par une flèche étiquetée $\mathbf{b}_j$ issue de l'origine $\mathbf{0}$. (La construction de ces indicateurs est discutée dans la Section 2.2, ci-dessous.) On peut donc parler des $\mathbf{a}_i$ et $\mathbf{b}_j$ comme, respectivement, indicateurs-lignes et indicateurs-colonnes pour $\mathbf{Y}$.

Les codes des mesures et des races sont (selon Jambu, 1977) :

$Y1$ longueur condylo-basale
$Y2$ longueur mâchoire supérieure
$Y3$ largeur bi-maxillaire
$Y4$ longueur carnassière supérieure
$Y5$ longueur de la première molaire supérieure
$Y6$ largeur première molaire supérieure

B bull-dog ; i chien-ind ; b berger allemand ; v lévrier ; c colley ; D doberman ; s setter ; y chien des Pyrénées ; r bas-rouge ; a briard ; M bull-mastif ; x boxer ; G dogue allemand ; g gronendal ; S saint-bernard ; p loup ; j canidé de Jussac.

La manière de représentation graphique par biplot est que chaque donnée centrée de $\mathbf{Y}$ est visualisée comme

$$y_{i,j} \longmapsto \mathbf{a}'_i \mathbf{b}_j \tag{1}$$

où $\mathbf{a}'_i \mathbf{b}_j$ est le produit scalaire de $\mathbf{a}_i$ et $\mathbf{b}_j$. Nous utilisons la même notation pour les indicateurs et pour les vecteurs définis par leurs coordonnées ainsi que pour la représentation graphique et l'approximation numérique. Donc, le symbole $\longmapsto$ signifie « est représenté par » aussi bien que « est approximé par ».

L'expression (1) signifie que la visualisation est opérée par le produit des longueurs de la flèche $\mathbf{b}_j$ et de la projection de $\mathbf{a}_j$ sur l'axe passant par $\mathbf{b}_j$, avec le signe positif ou négatif selon que la projection est dans la direction de $\mathbf{b}_j$ ou dans la direction opposée. La moyenne de chaque mesure est visualisée par l'origine.

Qualités d'ajustement : 0,9823 pour Y, 0,9998 pour YY', 0,5804 pour Y'Y

FIG 1. – Chiens et loups (données centrées) : biplot RMP avec métrique euclidienne.

Ces projections permettent de voir (Figure 1), entre autre, que la longueur condylo-basale (flèche $Y1$) est grande chez les dogues allemands (points étiquetés G) et petite chez les deux bull-dogs (points B). Pareillement, le colley (point c) a une faible largeur bimaxillaire (flèche $Y3$).

En plus d'approximer les données, un biplot approxime les dissimilarités entre les lignes de $\mathbf{Y}$, dites $diss$, par les distances entre leurs indicateurs, donc pour les lignes i et e

$$diss_{i,e} \longmapsto ||\mathbf{a}_i - \mathbf{a}_e||, \tag{2}$$

où $||\cdot||$ est la norme euclidienne. Dans cet article nous évitons les complications algébriques de l'approximation directe des dissimilarités par distances en utilisant le lien entre la matrice des dissimilaritées carrées et celle des « variances et covariances » $\mathbf{y}_i'\mathbf{y}_e$ des lignes $\mathbf{y}_1', \ldots, \mathbf{y}_n'$ de $\mathbf{Y}$, donc avec $\mathbf{YY}'$. Ce lien est

$$[[diss_{i,e}^2]] = \mathbf{11}'Diag(\mathbf{YY}') + Diag(\mathbf{YY}')\mathbf{11}' - 2\mathbf{YY}' \tag{3}$$

(Torgerson, 1958).

Nous appelons « forme » la matrice $\mathbf{YY}'$ en relation avec son application aux études des formes de crânes et autres objets (Goodall, 1991). Donc, nous discuterons l'approximation de la forme au lieu de celle des dissimilarités.

Pour les données centrées des chiens et des loups, les proximités et distances des indicateurs-crânes du biplot (Figure 1) indiquent les similarités et les différences entre crânes divers. On constate que la plupart des loups (p) ont des indicateurs sur la gauche, voisinant ceux des dogues (G) et du saint-bernard (S). Les indicateurs du reste des chiens sont à droite, ceux des bull-mastifs (M) et du boxer (x) étant les plus éloignés des loups, ce qui veut dire que se sont les races aux crânes les plus différents de ceux des loups. Les bergers allemands (b), par contre, ont des crânes plus similaires à ceux des loups.

Finalement, la configuration des flèches d'un biplot met en évidence la dispersion mesurée par la matrice de variances $\mathbf{YY}'$ (nous omettons les constantes dans les définitions de la variance et de la forme car elles ne changent rien dans la discussion de qualité proportionnelle d'approximation qui est étudiée dans cet article). Les longueurs des flèches et les cosinus des angles entre elles approximent, respectivement, les écarts-types et les corrélations des colonnes $\mathbf{y}_{(1)}, \ldots, \mathbf{y}_{(m)}$ de $\mathbf{Y}$, comme suit,

$$\sqrt{\mathrm{var}(\mathbf{y}_{(j)})} \longmapsto ||\mathbf{b}_j|| \tag{4}$$

et

$$corr(\mathbf{y}_{(j)}, \mathbf{y}_{(g)}) \longmapsto \cos(\mathbf{b}_j, \mathbf{b}_g). \tag{5}$$

Selon le biplot des crânes (Figure 1) il apparaît que la mesure $Y3$ (largeur bi-maxillaire) a la plus grande variabilité, et les mesures $Y1$ et $Y2$ (mesures de longueur du crâne) ont des variabilités moins grandes, tandis que le reste des mesures ont des variabilités si petites qu'on n'aperçoit pas leurs flèches. L'angle à peu près droit entre les flèches de $Y3$ et de $Y1$ indique une corrélation très faible entre ces mesures. L'angle aigu entre les flèches de $Y2$ et $Y1$ est évidence d'une corrélation positive entre ces deux mesures de longueur, tandis que l'angle obtus entre les flèches de $Y2$ et $Y3$ indique une corrélation négative avec la mesure de largeur.

La direction de la plus claire discrimination entre loups et chiens est plus ou moins celle de la flèche de $Y1$, donc la plus grande différence entre ces deux espèces est une longueur de crâne. Le rôle de la mesure de largeur $Y3$ est moins clair, mais le biplot suggère que les dogues, le saint-bernard, les bull-mastifs et les bull-dogs ont des crânes larges tandis que le colley et les lévriers ont des crânes plus étroits.

Donc, on a montré que le biplot représente les similarités et différences entre individus (crânes dans cet exemple), la dispersion des variables (mesures dans le cas présent), ainsi que les correspondances entre individus et variables (observations des mesures sur les crânes). On remarque que les représentations séparées des $\mathbf{a}_i$ et des $\mathbf{b}_j$ sont très courantes (voir, par exemple, Saporta, 1990, Fig. 8.9 et Fig. 8.10) mais que leur représentation simultanée (effectivement en biplot) est plus rare. Le biplot a l'avantage sur la plupart des graphiques de représenter les trois aspects des données et les relations entre eux, tandis que chaque autre graphique ne représente qu'un seul aspect. (Voir Gabriel, 1971, 1981, 1982; Gower and Hand, 1996, pour les caractéristiques du biplot.)

Il faut noter que la visualisation des données par biplot utilise seulement les indicateurs, en regardant leurs longueurs, distances, angles et projections mutuelles, mais n'utilise pas les axes. Le rôle de ces derniers est seulement dans la construction du biplot. C'est là une différence avec l'abord habituel de l'Analyse en Composantes Principales (ACP) qui essaye d'interpréter les axes. La visualisation par biplot se fixe directement sur les données elles-mêmes, les individus et les variables, n'utilisant les composantes principales que comme moyen possible du calcul de l'approximation. C'est un abord de présentation plutôt que d'analyse.

2.2. Calculs

En forme matricielle, on a écrit $\mathbf{Y}\boxed{(n \times m)}$ pour la matrice des données centrées sur les moyennes de chaque mesure, et on peut aussi écrire $\mathbf{A}\boxed{(n \times 2)}$ pour la matrice dont les lignes sont $\mathbf{a}'_i$, $i = 1, \ldots, n$, et $\mathbf{B}\boxed{(m \times 2)}$ pour la matrice dont les lignes sont $\mathbf{b}'_j$, $j = 1, \ldots, m$. La visualisation (1) des donnés par biplot ce fait donc au moyen de l'approximation matricielle

$$\mathbf{Y} \longmapsto \mathbf{AB}', \tag{6}$$

ce qui est équivalent à l'ajustement en rang 2 de $\mathbf{Y}$. Il s'agit de la collection d'approximations vectorielles des lignes de $\mathbf{Y}$

$$\mathbf{y}'_i \longmapsto \mathbf{a}'_i\mathbf{B}', \quad i = 1, \ldots, n \tag{7}$$

que Gower et Hand (1996, p.12) appellent «*predictions*» des données multidimensionnelles par le biplot.

Une construction des indicateurs,des lignes et des colonnes se fait au moyen de la méthode des moindres carrés par la solution de

$$\min_{\mathrm{H}\, de\, rang\, 2} ||\mathbf{W}^{1/2}(\mathbf{Y} - \mathbf{H})\mathbf{M}^{1/2}|| = \min_{\mathrm{A,B}\, de\, 2\, colonnes} ||\mathbf{W}^{1/2}(\mathbf{Y} - \mathbf{AB}')\mathbf{M}^{1/2}|| \tag{8}$$

(Gabriel, 1978a). La matrice $\mathbf{W}\boxed{(n \times n)}$ est une matrice diagonale dont les éléments diagonaux sont des pondérations des lignes de $\mathbf{Y}$ et $\mathbf{W}^{1/2}$ satisfait $\mathbf{W}^{1/2}\mathbf{W}^{1/2} = \mathbf{W}$, tandis que la matrice $\mathbf{M}\boxed{(m \times m)}$ est une métrique (semi définie positive) et $\mathbf{M}^{1/2}$ satisfait $\mathbf{M}^{1/2}\mathbf{M}^{1/2} = \mathbf{M}$. (Ces idées et les techniques qui en résultent peuvent être généralisées pour des pondérations par matrice $\mathbf{W}\boxed{(n \times n)}$ non négative quelconque et pour $\mathbf{W}^{1/2}\boxed{(n \times n)}$ symétrique qui satisfait $\mathbf{W}^{1/2}\mathbf{W}^{1/2} = \mathbf{W}$; Caussinus, 1986). Pour les biplots de la section présente on utilise $\mathbf{W} = \mathbf{I}_n$ et $\mathbf{M} = \mathbf{I}_m$, donc l'approximation est par moindres carrés simples, c'est-à-dire non pondérés et avec une métrique euclidienne canonique.

On remarque que ce minimum a des analogies en régression pour les cas où l'un des facteurs $\mathbf{A}$ ou $\mathbf{B}$ est connu (voir la Section 3.9 ci-dessous). Si $\mathbf{A}$ est connue (par exemple si les $\mathbf{a}'_i$s sont définis comme coordonnées des endroits où se trouvent les individus $i = 1, \ldots, n$) le minimum de (8) devient

$\min_{\mathbf{B}} ||\mathbf{W}^{1/2}(\mathbf{Y} - \mathbf{AB}')\mathbf{M}^{1/2}||$, donc celui d'une régression pondérée qui admet la solution

$$\mathbf{B}' = (\mathbf{A}'\mathbf{WA})^{-1}\mathbf{A}'\mathbf{WY}. \tag{9}$$

L'approximation $\mathbf{AB}'$ sera alors obtenue au moyen de la projection oblique $\mathbf{A}(\mathbf{A}'\mathbf{WA})^{-1}\mathbf{A}'\mathbf{WY}$ sur l'espace bidimensionnel des colonnes de $\mathbf{A}$ perpendiculairement à l'espace généré par $\mathbf{WA}$. On remarque que ce calcul ne dépend pas de la matrice $\mathbf{M}$.

Pareillement, si $\mathbf{B}$ est connue (par exemple, si les $\mathbf{b}'_j$ sont définis comme coordonnées des variables $j = 1, \ldots, m$ obtenues par analyse factorielle) le minimum de (8) devient $\min_{A} ||\mathbf{W}^{1/2}(\mathbf{Y} - \mathbf{AB}')\mathbf{M}^{1/2}||$ correspondant à une autre régression pondérée, avec la solution

$$\mathbf{A}' = (\mathbf{B}'\mathbf{MB})^{-1}\mathbf{B}'\mathbf{MY}'. \tag{10}$$

Dans ce cas, l'approximation $\mathbf{AB}'$ sera obtenue comme transposée de la projection oblique $\mathbf{B}(\mathbf{B}'\mathbf{MB})^{-1}\mathbf{B}'\mathbf{MY}'$ des lignes de $\mathbf{Y}$ sur l'espace bidimensionnel des colonnes de $\mathbf{B}$ perpendiculairement à l'espace généré par $\mathbf{MB}$. (Gower et Hand, 1996, p. 12, appellent ce calcul «interpolation»). Ici on remarque que ce calcul ne dépend pas de la matrice $\mathbf{W}$.

Dans la discussion présente on ne présume la connaissance ni de $\mathbf{A}$ ni de $\mathbf{B}$, mais on peut estimer les deux par itération entre les solutions ci-dessus. On commence par une approximation initiale de $\mathbf{A}$ avec laquelle on utilise (9) pour la première approximation de $\mathbf{B}$, puis on utilise (10) avec ce $\mathbf{B}$ pour la deuxième approximation de $\mathbf{A}$, etc., jusqu'à la convergence : cela s'appelle la méthode « *criss-cross regression* » (Wold, 1966 ; voir aussi Gabriel et Zamir, 1979). Effectivement, cette méthode consiste en itérations entre projections des lignes et des colonnes de $\mathbf{Y}$ sur deux espaces bidimensionnels, les coordonnées de chaque projection déterminant la base de l'espace pour la prochaine projection.

Ces calculs sont équivalents à ceux basés sur l'approximation en rang réduit selon le théorème de Householder et Young (1938). Donc, cette approximation est égale au produit

$$\mathbf{AB}' = \mathbf{W}^{-1/2}\sum_{k=1}^{2}\mathbf{u}_{(k)}\mathrm{d}_k\mathbf{v}'_{(k)}\mathbf{M}^{-1/2} \tag{11}$$

où

$$\mathbf{W}^{1/2}\mathbf{YM}^{1/2} = \sum_{k=1}^{rang(\mathrm{Y})}\mathbf{u}_{(k)}\mathrm{d}_k\mathbf{v}'_{(k)} \tag{12}$$

est la décomposition singulière avec $\mathbf{u}'_{(k)}\mathbf{u}_{(k')} = \mathbf{v}'_{(k)}\mathbf{v}_{(k')} = \delta_{k,k'}$ (symbole de Kronecker), et $\mathrm{d}_1 \geqslant \mathrm{d}_2 \geqslant \ldots \geqslant \mathrm{d}_{rang(\mathrm{Y})} > 0$. Mais, tandis que la décomposition singulière n'existe pas pour des matrices avec éléments manquants, la méthode « *criss-cross* » pondérée individuellement peut quelquefois être aussi adaptée à ce cas (Gabriel, 2003a).

Pour obtenir les indicateurs on peut définir

$$\mathbf{A}_{\{\lambda\}} = \mathbf{W}^{-1/2}(\mathbf{u}_{(1)}\mathrm{d}_1^{\lambda}, \mathbf{u}_{(2)}\mathrm{d}_2^{\lambda}) \tag{13}$$

et

$$\mathbf{B}_{\{\mu\}} = \mathbf{M}^{-1/2}(\mathbf{v}_{(1)}\mathrm{d}_1^{\mu}, \mathbf{v}_{(2)}\mathrm{d}_2^{\mu}). \tag{14}$$

Les représentations optimales au sens des moindres carrés (8) sont les

$$\mathbf{Y} \longmapsto \mathbf{A}_{\{\lambda\}}\mathbf{B}'_{\{\mu\}} \tag{15}$$

avec des λ et μ qui satisfont $\lambda + \mu = 1$; c'est-à-dire que

$$\mathbf{Y} \longmapsto \mathbf{A}_{\{\lambda\}}\mathbf{B}'_{\{1-\lambda\}} \tag{16}$$

est optimale pour λ quelconque. Dans certaines situations on pourra préférer des modalités de biplots où λ et μ ne satisfont pas la condition d'optimalité $\lambda + \mu = 1$ (voir la discussion dans la Section 5, ci-dessous.)

Deux autres calculs conduisant aux mêmes résultats sont les suivants, à partir des définitions généralisées de la forme et de la variance comme, respectivement, $\mathbf{YMY}'$ et $\mathbf{Y}'\mathbf{WY}$. Pour $\mathbf{YMY}'$, ou pareillement pour tout autre estimateur de la forme, le premier calcul obtient le minimum $\min\limits_{\mathrm{A}\,de\,rang\,2} ||\mathbf{W}^{1/2}(\mathbf{YMY}' - \mathbf{AA}')\mathbf{W}^{1/2}||$ au moyen de la décomposition spectrale

$$\mathbf{W}^{1/2}\mathbf{YMY}'\mathbf{W}^{1/2} = \sum_{k=1}^{rang(\mathrm{Y})} \mathbf{u}_{(k)}\mathrm{d}_k^2\mathbf{u}'_{(k)} \tag{17}$$

et donne l'approximation des moindres carrés

$$\mathbf{YMY}' \longmapsto \mathbf{W}^{-1/2}\sum_{k=1}^{2} \mathbf{u}_{(k)}\mathrm{d}_k^2\mathbf{u}'_{(k)}\mathbf{W}^{-1/2} = \mathbf{A}_{\{1\}}\mathbf{A}'_{\{1\}} \tag{18}$$

par le $\mathbf{A}_{\{\lambda\}}$ de (13) avec $\lambda = 1$. D'autres approximations seront

$$\mathbf{YMY}' \longmapsto \mathbf{A}_{\{\lambda\}}\mathbf{A}'_{\{\lambda\}} \tag{19}$$

pour d'autres valeurs de λ. On obtient aussi une approximation de $\mathbf{Y}$ par (15) au moyen de la régression (9) ce qui donne, comme en (14),

$$\mathbf{B}_{\{1-\lambda\}} = \mathbf{M}^{-1/2}(\mathbf{v}_{(1)}\mathrm{d}_1^{1-\lambda}, \mathbf{v}_{(2)}\mathrm{d}_2^{1-\lambda}). \tag{20}$$

Pour les moindres carrés on approxime la forme par (18) et les données par $\mathbf{A}_{\{1\}}$ et

$$\mathbf{B}_{\{0\}} = \mathbf{M}^{-1/2}(\mathbf{v}_{(1)}, \mathbf{v}_{(2)}) \tag{21}$$

ce qui est appelé biplot *RMP* (*Row Metric Preserving*).

Le deuxième calcul est pour la variance $\mathbf{Y}'\mathbf{WY}$, ou pareillement pour tout autre estimateur de la variance ; on cherche le minimum

$\min_{\mathbf{B}\, de\, rang\, 2} ||\mathbf{M}^{1/2}(\mathbf{Y}'\mathbf{W}\mathbf{Y} - \mathbf{B}\mathbf{B}')\mathbf{M}^{1/2}||$ au moyen de la décomposition spectrale

$$\mathbf{M}^{1/2}\mathbf{Y}'\mathbf{W}\mathbf{Y}\mathbf{M}^{1/2} = \sum_{k=1}^{rang(\mathrm{Y})} \mathbf{v}_{(k)}\mathrm{d}_k^2\mathbf{v}'_{(k)}, \tag{22}$$

donc les d_k^2 sont les valeurs propres de $\mathbf{Y}'\mathbf{W}\mathbf{Y}\mathbf{M}$ et les $\mathbf{M}^{-1/2}\mathbf{v}_{(k)}$ ses vecteurs propres. Par le théorème de Householder et Young on obtient l'approximation des moindres carrés

$$\mathbf{Y}'\mathbf{W}\mathbf{Y} \longmapsto \mathbf{M}^{-1/2}\sum_{k=1}^{2}\mathbf{v}_{(k)}\mathrm{d}_k^2\mathbf{v}'_{(k)}\mathbf{M}^{-1/2} = \mathbf{B}_{\{1\}}\mathbf{B}'_{\{1\}} \tag{23}$$

par le $\mathbf{B}_{\{\mu\}}$ de (14) avec $\mu = 1$. D'autres approximations seront

$$\mathbf{Y}'\mathbf{W}\mathbf{Y} \longmapsto \mathbf{B}_{\{\mu\}}\mathbf{B}'_{\{\mu\}} \tag{24}$$

pour d'autres valeurs de μ. Pour obtenir aussi une approximation de $\mathbf{Y}$ par (15) on se sert de la régression (10) et on trouve, comme en (13),

$$\mathbf{A}_{\{1-\mu\}} = \mathbf{W}^{-1/2}(\mathbf{u}_{(1)}\mathrm{d}_1^{1-\mu}, \mathbf{u}_{(2)}\mathrm{d}_2^{1-\mu}). \tag{25}$$

Pour les moindres carrés on approxime la variance par (23) et les données par $\mathbf{B}_{\{1\}}$ et

$$\mathbf{A}_{\{0\}} = \mathbf{W}^{-1/2}(\mathbf{u}_{(1)}, \mathbf{u}_{(2)}), \tag{26}$$

pour ce qui est appelé biplot *CMP* (*Column Metric Preserving*).

Les calculs de (23) et (26) sont connus comme étant l'*ACP* quand la matrice $\mathbf{Y}$ a des colonnes centrées ; les composantes principales et axes principaux sont les colonnes de $\mathbf{A}_{\{0\}}$ et de $\mathbf{B}_{\{1\}}$, respectivement.

Cet article discute les approximations par moindres carrés avec pondérations et métriques fixes, mais les mêmes principes sont encore valables pour les moindres carrés itératifs, donc pour les modèles bilinéaires généralisés (Falguerolles et Francis, 1992 ; van Eeuwijck, 1995 ; Gabriel, 1998). De plus, on peut généraliser la discussion aux approximations robustes (Gabriel et Odoroff, 1984 ; Daigle et Rivest, 1992 ; Ruiz-Gazen, 1996, Verboon, 1994) ou se servir du « *criss-cross regression* » pour ajuster le biplot aux données (Gabriel et Zamir, 1979 ; Gabriel, 1998).

2.3. Les qualités des approximations

On peut évaluer les qualités d'approximation séparément pour l'ajustement de $\mathbf{Y}$ par $\mathbf{A}_{\{\lambda\}}\mathbf{B}'_{\{\mu\}}$, de $\mathbf{Y}\mathbf{M}\mathbf{Y}'$ par $\mathbf{A}_{\{\lambda\}}\mathbf{A}'_{\{\lambda\}}$ et de $\mathbf{Y}'\mathbf{W}\mathbf{Y}$ par $\mathbf{B}_{\{\mu\}}\mathbf{B}'_{\{\mu\}}$. Dans cet article nous évaluons ces qualités pour une matrice centrée $\mathbf{X}$ quelconque (aux colonnes centrées) par le carré de la corrélation entre les éléments de $\mathbf{X}$ et les éléments de la matrice $\widehat{\mathbf{X}}$ qui l'ajuste, donc par

$$corr^2(\widehat{\mathbf{X}}, \mathbf{X}) := 1 - \frac{\min_{\theta} ||\mathbf{X} - \theta\widehat{\mathbf{X}}||^2}{||\mathbf{X}||^2} = \frac{\mathrm{tr}^2\{\mathbf{X}'\widehat{\mathbf{X}}\}}{\mathrm{tr}\{\mathbf{X}'\mathbf{X}\}\mathrm{tr}\{\widehat{\mathbf{X}}'\widehat{\mathbf{X}}\}} = \cos^2(\widehat{\mathbf{X}}, \mathbf{X}), \tag{27}$$

ce qui correspond dans certains cas au *RV* d'Escoufier (1973). Nous avons choisi ce coefficient qui évalue l'ajustement *proportionnel*, plutôt que l'ajustement absolu, parce qu'on se sert de graphiques pour comparer plutôt que pour mesurer.

On pourrait aussi utiliser d'autres coefficients (Ramsay *et al.*, 1984), comme celui de Gower (1971) ou de Lingoes et Schöneman (1974), mais cela mène aux mêmes conclusions (Gabriel, 2002).

2.4. Un autre biplot des chiens et loups

Les qualités d'approximation du biplot (Figure 1) des données centrées des chiens et loups sont 98,23 % pour **Y**, 99,98 % pour **YY′** et 58,04 % pour **Y′Y**. Le coefficient très élevé pour **Y** est dû à une approximation excellente des trois variables de grande dispersion et cache la pauvre qualité d'approximation des autres variables qui ont des dispersions plus petites. On peut tenter d'égaliser les approximations de toutes les variables en représentant les données réduites, c'est-à-dire après division des observations centrées de chaque variable par leurs écarts-types.

Qualités d'ajustement : 0,8304 pour **Y**, 0,9731 pour **YY'**, 0,6868 pour **Y'Y**

FIG 2. – Chiens et loups (données standardisées) : biplot *RMP* avec métrique euclidienne.

Le biplot des données réduites (Figure 2) a des qualités d'approximation de 83,04 % pour $\mathbf{Y}$, 97,31 % pour $\mathbf{YY}'$ et 68,68 % pour $\mathbf{Y}'\mathbf{Y}$. Ces qualités sont plus faibles que celles obtenues pour les données non réduites, mais elles concernent toutes les variables et tentent de les ajuster également. Les flèches de toutes les six mesures ont des longueurs proches de 1, ce qui n'est pas étonnant car tous les écarts-types ont été standardisés à 1. La proximité des longueurs à 1 indique la qualité des approximations des variables : la longue flèche de $Y3$ montre que cette variable est moins bien approximée que les autres.

On remarque la faible qualité de l'approximation de la variance $\mathbf{Y}'\mathbf{Y}$ par les biplots des Figures 1 et 2, en comparaison des excellentes qualités d'approximation des données Y et de la forme $\mathbf{YY}'$ (voir la discussion dans la Section 5 ci-dessous). C'est caractéristique du mode RMP choisi pour ces biplots et pour les autres de la Section 3 ci-dessous, mode que nous avons choisi pour que nos biplots soient comparables aux graphiques des ACP publiés par d'autres auteurs.

En regardant ce biplot (Figure 2) on aperçoit trois directions de variation des mesures. La première est celle d'une gerbe de flèches de mesures dentaires ($Y4$, $Y5$ et $Y6$), la deuxième celle de la mesure de largeur du crâne ($Y3$), et la troisième celle des mesures de longueur du crâne ($Y1$ et $Y2$). Ces dernières sont positivement corrélées avec les mesures dentaires et négativement avec la largeur. On constate que la plupart des loups, ainsi que les dogues et le saint-bernard, ont les indicateurs dans la direction générale de la gerbe ($Y4$, $Y5$ et $Y6$), ce qui indique qu'ils ont des grandes dents, tandis que le reste des chiens ont des plus petites dents. En regardant les projections des indicateurs-crânes sur l'axe de $Y3$ on observe aussi que les bull-dogs, le bull-mastif et le saint-bernard ont des crânes larges alors que les lévriers et le colley, entre autres, ont des crânes étroits. Il faut noter que Jambu (1977, Fig. 4) a effectivement produit un biplot de ces données et s'en est servi pour identifier les variables qui discriminent entre les chiens et les loups et celles qui distinguent les animaux aux proportions de museaux différentes. Cet auteur est arrivé à ces interprétations en observant les projections des indicateurs-crânes sur les axes principaux, tandis que, au lieu d'interpréter ces axes, nous préférons l'interprétation plus élémentaire des projections sur les axes des indicateurs des mesures elles-mêmes.

3. Le rôle des métriques

3.1. La métrique choisie selon les paradigmes statistiques

La Section 2 a illustré l'avantage d'une représentation des données réduites en comparaison avec la représentation des données seulement centrées. Nous avons utilisé l'ajustement par moindres carrés simples parce qu'il optimise la représentation de la variabilité des données, soit en forme centrée, soit réduite. Parfois, au lieu de la meilleure représentation de la variabilité, ce qu'on appellera «l'abord classique», il est d'intérêt de souligner certaines sources de variabilité à la place d'autres sources : cela peut se faire en utilisant

la méthode des moindres carrés avec une métrique qui réduit l'importance des sources considérées moins intéressantes.

Selon les paradigmes de la statistique inférentielle, les pondérations $\mathbf{W}$ et la métrique $\mathbf{M}$ sont définies par le modèle probabiliste présumé pour les observations. Par exemple, pour une matrice $\mathbf{Y}$ dont les lignes sont les moyennes de n échantillons sur m variables, la diagonale de $\mathbf{W}$ contient les effectifs des échantillons tandis que $\mathbf{M}^{-1}$ est l'estimateur de la variance « intra » des échantillons. Par contre, les paradigmes de l'Analyse des Données présument des pondérations égales $\mathbf{W} = \mathbf{I}_n$ et choisissent une métrique $\mathbf{M}$ qui concentre la représentation sur certains aspects d'intérêt. Cela se fait au moyen de la standardisation des colonnes par la variabilité engendrée en $\mathbf{M}^{-1}$, ce qui met en évidence le reste de la variabilité. Finalement, le paradigme des projections révélatrices mène aux métriques définies au moyen des données elles-mêmes par des expressions du type

$$\mathbf{M} = (\mathbf{Y}'\mathbf{K}\mathbf{Y})^{-1}, \tag{28}$$

où $\mathbf{K}\boxed{(n \times n)}$ est choisie pour accentuer certains aspects de la dispersion des données. Pour ces métriques la représentation en biplot est invariante par transformations affines, donc le problème du choix des échelles, qui a souvent préoccupé les utilisateurs de l'ACP, disparaît (Caussinus, 1992).

On remarque que le choix de métrique s'impose seulement sur l'analyse des données, et non sur la représentation graphique qui doit toujours être euclidienne parce que c'est à cela que l'œil est accoutumé. D'autre part, la représentation par produits scalaires permet une flexibilité de choix d'échelles dont on peut profiter pour mieux visualiser les données. En effet l'approximation

$$\mathbf{Y} \longmapsto r\mathbf{A}(r^{-1}\mathbf{B})', \tag{29}$$

donc

$$y_{i,j} \longmapsto (r\mathbf{a}_i)' \left(\frac{1}{r}\ \mathbf{b}_j\right) \quad \forall i, j, \tag{30}$$

est égale à celle de (6) et (1), quel que soit le r ($r \neq 0$) choisi (voir la discussion plus générale dans la Section 5 ci-dessous). Pour les graphiques il est bon d'ajuster les facteurs calculés et de présenter $r\mathbf{A}$ et $r^{-1}\mathbf{B}$ avec r choisi de sorte que les lignes de chaque facteur ajusté couvrent la région du graphique aussi bien que possible (c'est particulièrement important si n et m sont très différents). Si, par exemple, les $\mathbf{a}_i$ sont à peu près deux fois plus grands que les $\mathbf{b}_j$ on utilisera $r = 1/2$. Des ajustements de ce type ont été faits pour tous les biplots présentés ici.

Les exemples suivants illustrent ces rôles de diverses métriques.

3.2. Une métrique de contiguïté

On pourrait se servir d'une mesure g de la contiguïté entre toutes paires d'individus, de manière que $g_{i,e} = g_{e,i}$ exprime la contiguïté des individus i et e. Dans l'esprit de l'analyse de contiguïté (Lebart, 1969) on utilise donc (28)

comme métrique, où

$$\mathbf{K} = (Diag_{[g]} - \mathbf{G}) / \sum g, \tag{31}$$

quand $\mathbf{G} = (g_{i,e}, i, e = 1, \ldots, n)$ est la matrice des contiguïtés, $\mathbf{g}$ le vecteur des sommes des lignes de $\mathbf{G}$, et $\sum g$ la somme de ses éléments. Des exemples de contiguïté sont les similarités entre espèces ou races, la proximité géographique des individus observés (Lebart, 1969) ou bien la similarité des structures géologiques d'endroits divers (Faraj, 1993, 1994). Les représentations qui en résultent suppriment la variation contiguë et soulignent la variation non contiguë.

Pour l'exemple présent, on pourrait définir les contiguïtés comme suit,

$g_{i,e} = g_{e,i} = 0$ si i est un chien et e un loup,
$g_{i,e} = g_{e,i} = 1/3$ si i et e sont deux chiens de races différentes,
$g_{i,e} = g_{e,i} = 2/3$ si i et e sont deux chiens de la même race ou deux loups,
$g_{i,i} = 1$ si i est un animal quelconque.

(On remarque que cet exemple étend un peu la définition de Lebart en permettant des contiguïtés intermédiaires entre 0 et 1.)

Qualités d'ajustement : 0,4088 pour **Y**, 0,2517 pour **YY'**, 0,6474 pour **Y'Y**

FIG 3. – Chiens et loups (données standardisées) : biplot RMP avec métrique de contiguïté des espèces et des races.

L'ajustement avec cette métrique réduit les différences intra-espèces et intra-races et par conséquent souligne les différences d'une espèce ou d'une race à une autre. Les qualités d'approximation du biplot qui en résulte (Figure 3) sont beaucoup plus faibles que celles du biplot avec métrique euclidienne, parce que le biplot présent sert à éclairer certains aspects des données plutôt qu'à décrire leur variation totale. Il montre une séparation complète des indicateurs des deux espèces qui est manifestée au maximum dans la direction horizontale, donc par les mesures de longueur dentaire. Cependant, ce biplot ne donne aucune information notable sur les races diverses des chiens. D'autre part, il indique qu'il y a deux loups avec des observations extrêmes, apparemment sur $Y6$, une largeur de molaire.

3.3. Une contiguïté définie par les données elles-mêmes

Un autre abord intéressant est d'obtenir la métrique à partir des données elles-mêmes. On peut le faire en définissant $g_{i,e} = g_{e,i} = 1$ ou 0 si un coefficient de dissimilarité entre observations i et e est petit ou grand, respectivement. En supprimant la variabilité entre individus similaires ce biplot accentue les «traits structuraux» ou «globaux» (Faraj, 1993; Lebart, 2001). Une variante de cette idée est de choisir la métrique directement au moyen de certaines fonctions de dissimilarité statistique, comme discuté ci-dessous pour les métriques S et T.

3.4. La métrique S des observations centrales

La métrique préconisée par Caussinus et Ruiz-Gazen (1993) pour souligner les outliers multidimensionnels sera appelée S(β_1), où β_1 est une constante petite, de l'ordre de 0,05 ou 0,10. Sa matrice est

$$\mathbf{M} = \left[\sum_{i=1}^{n} K_i (\mathbf{y}_i - \overline{\mathbf{y}})(\mathbf{y}_i - \overline{\mathbf{y}})' \Big/ \sum_{i=1}^{n} K_i \right]^{-1}, \tag{32}$$

où $\mathbf{y}'_i$ est la i-ième ligne de $\mathbf{Y}$, $\overline{\mathbf{y}} = \sum_{i=1}^{n} \mathbf{y}_i/n$, $\mathbf{C} = \sum_{i=1}^{n} (\mathbf{y}_i - \overline{\mathbf{y}})(\mathbf{y}_i - \overline{\mathbf{y}})'/n$, $||\mathbf{y}_i - \overline{\mathbf{y}}||^2_{\mathbf{C}^{-1}} = (\mathbf{y}_i - \overline{\mathbf{y}})'\mathbf{C}^{-1}(\mathbf{y}_i - \overline{\mathbf{y}})$ et $K_i = \exp\{-\beta_1 ||\mathbf{y}_i - \overline{\mathbf{y}}||^2_{\mathbf{C}^{-1}}/2\}$. Son inverse $\mathbf{M}^{-1}$ est un estimateur robuste de la variance avec des pondérations importantes pour les lignes $\mathbf{y}'_i$ qui sont proches de la ligne moyenne $\overline{\mathbf{y}}'$. Par conséquent, le biplot avec cette métrique souligne l'effet des lignes éloignées de la moyenne, et spécialement des outliers.

Le biplot des données des chiens et des loups ajusté par la métrique S(0,05) sert à souligner l'aspect des données éloignées du centroïde et par conséquent ses qualités d'approximations générales sont beaucoup plus faibles que celles du biplot avec métrique euclidienne. Ce biplot (Figure 4) révèle trois outliers, un bull-dog (#1 de la liste de Jambu, 1977) avec une très grande valeur de $Y3$, un loup (#39 de la liste) avec une très grande valeur de $Y6$ et un autre loup (#34 de la liste) dont la direction du centroïde (à l'origine du biplot) est à peu près celle des combinaisons linéaires $Y4$-$Y6$, $Y5$-$Y6$ ou $0,5(Y4 + Y5)$-$Y6$ qui

Qualités d'ajustement : 0,3098 pour **Y**, 0,0981 pour **YY'**, 0,5494 pour **Y'Y**

FIG. 4. – Chiens et loups : biplot RMP avec métrique S(0, 5).

peuvent être construites sur le biplot par les méthodes usuelles de construction vectorielle. (La justification de pareilles constructions sur le biplot est que toute combinaison linéaire de colonnes de **Y** est représentée sur le biplot par la même combinaison linéaire des indicateurs-colonnes (Gabriel, 1995a, Section 9.3.1)). On peut vérifier que ces conclusions sont justes en identifiant ces crânes sur les nuages de toutes les paires de mesures (Figure 5). On note que les deux premiers outliers ont été soulignés aussi par le biplot avec métrique de contiguïté (Figure 3). (Caussinus, Hakam et Ruiz-Gazen (2001) ont tracé en leur Figure 8 exactement le même nuage d'indicateurs-crânes avec métrique S(0,05) et ont identifié les mêmes outliers. Comme ils n'ont pas ajouté les indicateurs-mesures à leur graphique, ils n'ont pas pu visualiser les variables sur lesquelles les outliers étaient extrêmes bien qu'ils les aient indiquées au moyen de leurs corrélations avec les axes principaux.)

En comparant l'utilisation des métriques de contiguïté (Figure 3) et de voisinage à la moyenne (Figure 4) il faut prendre en compte que la première a utilisé l'information sur l'espèce et la race de chaque crâne, tandis que la dernière a seulement utilisé l'information des donnés **Y**. Cette dernière est donc l'outil préféré pour souligner les outliers quand on n'a pas d'information additionnelle ; la première sera préférée si l'on connaît les contiguïtés et veut souligner des effets non contigus. Une possibilité intéressante serait d'utiliser

FIG 5. – Chiens et loups (standardisées) : nuages de toutes les paires de mesures (outliers #1, #34 et #39 étiquetés).

la moyenne de ces deux métriques pour souligner séparément les outliers de chaque espèce. Le biplot qui en résulte (Figure 6) est presque le même que celui construit avec la métrique de contiguïté, donc il ne révèle rien de nouveau sur ces données. Cependant, il est bien possible qu'un tel mélange de métriques puisse parfois servir à révéler des outliers multidimensionnels de groupes contigus qui n'apparaissent pas comme outliers multidimensionnels de l'ensemble.

Bien entendu, ces méthodes sont préconisées pour la découverte d'outliers multidimensionnels, pas pour celle d'outliers sur variables individuelles.

Qualités d'ajustement : 0,4167 pour **Y**, 0,2459 pour **YY'**, 0,6980 pour **Y'Y**

FIG 6. – Chiens et loups (standardisées) : biplot *RMP* avec métrique « mélangée ».

3.5. L'exemple des insectes de Lubischew

Nous continuons la discussion avec un autre exemple : celui de six mesures faites sur une collection de 74 insectes dont l'entomologiste avait pensé *a priori* qu'il y avait trois classes correspondant, respectivement, aux individus numérotés 1-21, 22-43, et 44-74 (Lubischew, 1962). Les mesures réduites sur les biplots utilisant $\mathbf{W} = \mathbf{1}_{74}$ et $\mathbf{M} = \mathbf{I}_6$, étaient :

TU largeur de la première articulation du premier tarsus (micron),

TD largeur de la deuxième articulation du premier tarsus (micron),

TM largeur maximale de l'aedeagus au front (micron),

FA angle frontal de l'aedaegus (en unités de 7,5 degrés),

WH largeur maximale de la tête (en unités de 0,01 mm),

SA largeur de l'aedeagus vu du côté (microns).

Les indicateurs-individus $\mathbf{a}_i$ du biplot *RMP* (Figure 7) montrent que les insectes ont une dispersion qui ressemble un peu à un fer à cheval. Si on n'avait pas eu l'idée *a priori* de trois groupes, on ne les aurait pas aperçus par cette visualisation, bien que la qualité d'ajustement de la forme soit très élevée (96,34 %). En effet, l'examen de ces indicateurs-individus par Caussinus (1992)

Qualités d'ajustement : 0,8000 pour **Y**, 0,9634 pour **YY'**, 0,8580 pour **Y'Y**

FIG. 7. – Insectes (données standardisées) : biplot RMP avec métrique euclidienne.

« ne donne qu'une pâle information sur la structure du nuage des individus » (Caussinus, Hakam et Ruiz-Gazen, 2002, section 5.2.1).

Le biplot n'est pas borné au nuage des indicateurs-individus examinés par Caussinus *et coll.*, mais y ajoute l'éventail des indicateurs-mesures, c'est-à-dire des flèches $\mathbf{b}_j$. En regardant les projections des $\mathbf{a}_i$ sur les $\mathbf{b}_j$, ce qui a la qualité d'ajustement de 80 %, on voit que les insectes à un bout du fer à cheval présentent des valeurs grandes sur SA et FA et faibles sur TU et WH, tandis que le contraire est vrai pour celles à l'autre bout du fer. Les insectes au centre du fer ont des grandes valeurs surtout sur TD, TM et FA.

En regardant les flèches on trouve, grosso modo, deux faisceaux. Les mesures TU et WH sont fortement corrélées ; les quatre autres mesures le sont moins fortement ; enfin, il existe peu de corrélation entre les deux groupes de mesures. Cette description est assez fiable, la qualité d'ajustement de la variance étant 85,80 %.

Cette présentation des données d'insectes, utilisant la métrique euclidienne, les a regardées en projection sur le plan de variabilité maximale, ce qui est une propriété de l'ACP classique. L'addition d'autres projections peut approfondir l'étude des données en utilisant des métriques qui servent à souligner des aspects divers d'intérêt particulier.

Qualités d'ajustement : 0,3050 pour Y, 0,1254 pour YY', 0,5845 pour Y'Y

FIG. 8. – Insectes : biplot RMP avec métrique S(0,05).

On a déjà vu (Section 3.4) que la métrique $S(\beta_1)$ de Caussinus et Ruiz-Gazen (1993) sert à souligner les outliers multidimensionnels. Le biplot des données d'insectes qui résulte du choix $\beta_1 = 0,05$ (Figure 8) accentue les individus rares et, par conséquent, a des qualités assez faibles d'ajustement pour l'ensemble (30,50 % pour les données, 12,54 % pour la forme et 58,45 % pour les variances). Ce graphique ne révèle aucun outlier, bien que les insectes numérotés 67 et 68 aient des valeurs très petites sur SA et FA, 22 et 41 des valeurs très grandes sur TD et TM. (Caussinus, Hakam et Ruiz-Gazen, 2002, remarquent que certains de ces individus « sont un peu périphériques » mais, comme leurs graphiques ne représentent pas les variables, ne discutent pas les mesures sur lesquels ces individus sont rares. D'autre part, il faut signaler que ces auteurs ajoutent un test concluant que la présence d'outliers n'est pas significative.)

3.6. La métrique T des observations voisines

Un aspect des données qui est souvent de grande importance est la variation globale, par contraste avec la variation locale qu'on peut estimer par la métrique $\mathrm{T}(\beta_2)$ de Caussinus et Ruiz (1990) dont la matrice $\mathbf{M}$ est définie par

$$\mathbf{M}^{-1} = \sum_{i=1}^{n}\sum_{e=1}^{n} L_{i,e}(\mathbf{y}_i - \mathbf{y}_e)(\mathbf{y}_i - \mathbf{y}_e)' \Big/ \sum_{i=1}^{n}\sum_{e=1}^{n} L_{i,e}, \tag{33}$$

où $L_{i,e} = \exp\left[-\beta_2 \|\mathbf{y}_i - \mathbf{y}_e\|^2_{\mathbf{C}^{-1}}/2\right]$. Cette matrice $\mathbf{M}^{-1}$ est un estimateur robuste de la covariance avec pondération importante des paires $\mathbf{y}'_i$, $\mathbf{y}'_e$ de lignes qui sont très similaires. Donc l'utilisation de cette métrique accentue les paires où $\mathbf{y}'_i$ est très différent de $\mathbf{y}'_e$. C'est-à-dire, cette métrique ignore les différences locales et il en résulte un « biplot type global ». Le paramètre β_2 détermine quelle partie des distances est accentuée; donc les plus petits β_2 considèrent comme globales les seules distances très grandes, tandis que les plus grands β_2 considèrent aussi comme globales des distances moins grandes.

Qualités d'ajustement : 0,7067 pour Y, 0,7551 pour YY', 0,8931 pour Y'Y

FIG. 9. – Insectes : biplot RMP avec métrique $\mathrm{T}(2,0)$.

Le biplot des insectes ajusté avec métrique T(2,0) (Figure 9) révèle trois groupes plutôt homogènes mais il y a un doute sur la façon d'affecter certains insectes (numérotés 4,6,10,16,17). Une version robuste de ce biplot

Qualités d'ajustement : 0,7155 pour Y, 0,7659 pour YY', 0,8911 pour Y'Y

FIG 10. – Insectes : biplot RMP robuste avec métrique $T(2,0)$.

(voir ci-dessous) produit une meilleure séparation entre les trois groupes, ne présentant plus de doutes sur l'affectation des divers individus (Figure 10). Il est remarquable que cette classification en trois groupes coïncide précisément avec les classes suggérées par l'entomologiste. En mettant les trois groupes de points en rapport avec les flèches on trouve que le premier groupe (individus 1-21) a de grandes valeurs sur toutes les mesures, tandis que les autres groupes diffèrent sur TM et SA, le deuxième groupe (insectes 22-43) y ayant des valeurs plus élevées, et sur TU et WH où le troisième groupe (insectes 44-74) a des valeurs plus élevées. Pour résumer, les insectes du premier groupe sont généralement grands, tandis que les deux autres groupes diffèrent sur une combinaison $(TM + SA)$-$(TU + WH)$ des variables. (Caussinus, Hakam et Ruiz-Gazen, 2003, arrivent aux mêmes conclusions sur la séparation des groupes mais, comme ils n'ont pas inclus les indicateurs-variables sur leur graphique, ils n'identifient pas les variables qui distinguent les groupes. D'autre part, leur test indique au moins deux dimensions significatives et justifie la conclusion qu'il y a une structuration en au moins trois groupes.)

La méthode robuste conduisant au dernier biplot utilise un estimateur robuste $\mathbf{S}_{robuste}$ de la matrice de variance dû à Ruiz-Gazen (1996). On se sert du calcul des indicateurs-colonnes en $\mathbf{B}_{\{0\}}$ par les valeurs et vecteurs propres de

$\mathbf{Y}'\mathbf{Y}\mathbf{S}_{robuste}^{-1}$ et celui des indicateurs-lignes par $\mathbf{A}_{\{1\}} = \mathbf{Y}\mathbf{S}_{robuste}^{-1}\mathbf{B}_{\{0\}}$, comme dans (23) et (26).

On a montré que l'emploi de la métrique $\mathrm{T}(\beta_2)$ de Caussinus et Ruiz a révélé les trois groupes sans utiliser d'information *a priori* sur ces groupes. Ce devrait donc devenir un outil important pour l'analyse de données.

3.7. La métrique de Mahalanobis pour les moyennes multidimensionnelles

D'autres représentations sont possibles si l'on utilise la classification *a priori* des insectes en trois groupes. On pourrait se servir d'une métrique de contiguïté avec $g(i,e) = 1$ si les individus i et e sont de la même classe, et $g(i,e) = 0$ s'ils appartiennent à des classes différentes. Il est bien connu que c'est équivalent à définir cette métrique comme l'inverse de la variance « intra » (Burtschy et Lebart, 1991). On arrive à un biplot (Figure 11) qui souligne les différences entre les classes connues *a priori*. On y trouve une différentiation de classes un peu plus accentuée que celle utilisant $\mathrm{T}(\beta_2)$. La dispersion dans chaque classe peut être visualisée sur le plan par une ellipse de concentration définie, pour les n_{g} indicateurs $\mathbf{a}_i$, de la classe g, par les vecteurs $\mathbf{a}$ qui satisfont

$$(\mathbf{a} - \overline{\mathbf{a}}_{\mathrm{g}})'\mathbf{S}_{\mathrm{g}}^{-1}(\mathbf{a} - \overline{\mathbf{a}}_{\mathrm{g}}) \leqslant k \tag{34}$$

où le centroïde et la variance de la classe sont respectivement

$$\overline{\mathbf{a}}_{\mathrm{g}} = \sum_{i \in \mathrm{g}} \mathbf{a}_i / n_{\mathrm{g}} \tag{35}$$

$$\mathbf{S}_{\mathrm{g}} = \sum_{i \in \mathrm{g}} (\mathbf{a}_i - \overline{\mathbf{a}}_{\mathrm{g}})(\mathbf{a}_i - \overline{\mathbf{a}}_{\mathrm{g}})' / n_{\mathrm{g}} \tag{36}$$

et k est la fractile 0,90 du chi-carré à 2 degrés de liberté, donc, pour une distribution binormale, devrait inclure 90 % des observations. On voit que la dispersion de la troisième classe est un petit peu plus grande que celles des autres classes. Enfin, aucune classe ne présente d'outliers.

En fait, si la classification est connue, on a l'alternative de représenter directement les moyennes $\overline{y}_{\mathrm{g},j}$, $j = 1, \ldots, m$, des classes $\mathrm{g} = 1, ..., \mathrm{G}$ (Figure 12) par des indicateurs-échantillons $\overline{\mathbf{a}}_{\mathrm{g}}$ de (35) qui sont les centroïdes des indicateurs-individus du biplot des données (Figure 11). Les indicateurs-variables $\mathbf{b}_j$ sont les mêmes dans les deux biplots et leur rôle est d'aider à l'interprétation des positions des autres indicateurs. Si, par exemple, on voit que l'indicateur d'une classe est dans la direction de l'indicateur d'une certaine variable, on peut conclure que cette classe a une moyenne élevée pour cette variable.

Les calculs pour ce biplot sont conformes à ceux d'une analyse de variance multidimensionnelle, donc on l'appelle biplot *MANOVA* (Gabriel, 1972, 1981, 1995b). La métrique de « Mahalanobis » qui convient à cette manière

Qualités d'ajustement : 0,6635 pour Y, 0,7064 pour YY', 0,6711 pour Y'Y

FIG 11. – Insectes (données standardisées) : biplot *RMP* avec métrique de contiguïté dans classes et ellipses de concentration des classes (pour inclure 90 % des individus d'une distribution normale).

de regarder les données est l'inverse de la matrice de covariances « intra »

$$\mathbf{S}_{\text{intra}} = \sum_{g=1}^{G} n_g \mathbf{S}_g / n. \tag{37}$$

La distance $||\bar{\mathbf{a}}_g - \bar{\mathbf{a}}_{g'}||$ entre deux indicateurs-échantillons $\bar{\mathbf{a}}_g$ et $\bar{\mathbf{a}}_{g'}$ du biplot *MANOVA* approxime la dissimilarité Mahalanobis entre les moyennes des classes g et g'. (Nous évitons l'expression « distance Mahalanobis » pour souligner la différence entre le concept statistique de dissimilarité et le concept géométrique de distance que nous utilisons sur le biplot.) On peut visualiser les tests T-carré entre paires de classes au moyen de cercles d'incertitude pour toutes les classes

$$||\mathbf{a} - \bar{\mathbf{a}}_g|| \leqslant \sqrt{\chi^2_{(G,0,95)}/n_g}\ \frac{\sqrt{1+r_g}}{1+\sqrt{r_g}} \tag{38}$$

où

$$r_g = \max\{n_g / \max n, \min n/n_g\}. \tag{39}$$

FIG. 12. – Insectes : biplot RMP des moyennes des classes avec métrique «intra» et cercles d'incertitude pour tests de signification au niveau de 5 % par paire.

Les différences entre les classes g et g′ sont considérées significatives à 5 % si les cercles d'incertitude autour de $\bar{\mathbf{a}}_g$ et $\bar{\mathbf{a}}_{g'}$ sont disjoints. Si les cercles se coupent, ces différences ne sont pas considérées significatives. Ces décisions basées sur la position des cercles sont approximativement les mêmes que celles des tests T-carré de Hotelling entre paires de classes (Gabriel, 1972, 1995b).

Sur le biplot $MANOVA$ des moyennes des échantillons des trois classes d'insectes (Figure 12) on aperçoit que les trois échantillons sont tous différents significativement car aucun des cercles ne se coupent. En relation avec les flèches on voit que le premier groupe (étiqueté a) a des niveaux élevés de toutes les six variables, le groupe b a des niveaux élevés de SA et TM, faibles de WH et TP, tandis que le groupe c a des niveaux élevés de TP et WH et des niveaux faibles des quatre autres variables.

Un exemple artificiel de l'utilisation des cercles d'incertitude est créé en divisant les observations de deux des classes aléatoirement en sous-échantillons. Ainsi, les différences entre les sous-échantillons de chaque groupe sont purement aléatoires. Sur le biplot $MANOVA$ de ces données subdivisées (Figure

FIG. 13. – Insectes : biplot *RMP* des moyennes de sous-échantillons avec métrique «intra» et cercles d'incertitude pour tests de signification au niveau de 5 % par paire.

13) on voit que les cercles des sous-échantillons de chaque classe se coupent, c'est-à-dire que les différences entre ces sous-échantillons ne sont pas significatives. D'autre part, la plupart des cercles des sous-échantillons de classes différentes sont disjoints, montrant des tests significatifs. L'exception est le plus petit échantillon de la classe b dont le cercle coupe celui de la classe a. Donc, pour ce petit échantillon (5 insectes), on arriverait à accepter l'égalité de a et b bien qu'elle ne soit pas vraie – une erreur de seconde espèce. Néanmoins, la plupart de ces tests graphiques donnent des conclusions correctes.

On remarque que les ellipses de concentration (Figure 11) et les cercles d'incertitude (Figure 12) sont bien différents. Les ellipses représentent la variabilité de chaque classe, tandis que le cercles sont basés sur la variabilité d'échantillonnage des moyennes des classes et leurs rayons sont inversement proportionnels aux racines carrées des effectifs des échantillons de ces classes.

3.8. Autres biplots

La représentation par biplot type $MANOVA$ a des analogies pour les tableaux de contingence où l'on peut construire des biplots pour les profils des lignes et des colonnes (Gabriel and Odoroff, 1990; Gabriel, Galindo et Vicente-Villardón, 1998) ou pour leurs transformations logarithmiques (Gabriel, 1995b). Une autre application de cette méthode concerne les données univariables rangées selon deux facteurs, quand on utilise une métrique de corrélation zéro et de variances égales (voir l'exemple de la Section suivante). Une utilisation importante de cette application univariable est l'étude de l'interaction des environnements et des génotypes (Gauch, 1992, Gauch et Zobel, 1997, Yan *et al.*, 2000).

Gower et Hand (1996, Ch. 5) appellent cette sorte d'application « *canonical biplots* » et y incluent les biplots construits pour l'analyse des corrélations canoniques (Haber, 1975; ter Braak, 1990). Il faut aussi mentionner les biplots de la régression en rang réduit (ter Braak, 1994) et le « biplot du coefficient de variation » (Underhill, 1990).

Des biplots interactifs ont été utilisés par Sparks, Adolphson et Phatak (1997) pour contrôler des processus industriels.

Une autre approche est fondée sur l'analogie entre les composantes principales de données normales et les fragilités multidimensionnelles de données de survie; on espère présenter des biplots représentant ces fragilités dans un article prochain.

3.9. Quasi-biplots

On appellera quasi-biplot une représentation graphique par produits scalaires d'indicateurs-lignes et indicateurs-colonnes dans laquelle une de ces collections d'indicateurs est donnée *a priori* et seule l'autre est calculée pour ajuster les données. Quand l'ajustement est opéré par moindres carrés l'autre collection d'indicateurs est calculée par régression comme discuté dans la Section 2.2, ci-dessus.

On rencontre un exemple où $\mathbf{A}$ est donnée quand les individus (lignes) sont (placés dans) des endroits i et on utilise les coordonnés géographiques de chaque endroit i comme $\mathbf{a}_i$ (Kempton, 1984). Le quasi-biplot qui résulte du calcul des $\mathbf{b}_j$ par (9) est une carte sur laquelle les individus sont représentés par leurs coordonnés géographiques $\mathbf{a}_i$ et les variables par des flèches $\mathbf{b}_j$ qui indiquent la répartition des variables sur la région des localisations (Gabriel, 1987, 1988). Pour des exemples à $\mathbf{B}$ donné on peut penser aux coordonnées $\mathbf{b}_j$ des variables obtenues préalablement par une méthode d'analyse factorielle, ou aux coordonnées $\mathbf{b}_j$ des catégories de variables qualitatives déjà obtenues par des analyses comme celle des correspondances multiples (Benzécri, 1978; Greenacre, 1988). Les $\mathbf{a}_i$ calculés par (10) vont alors permettre de représenter les approximations des valeurs des individus sur chaque variable ou sur chaque catégorie (Gabriel, 1995a).

Une approche intéressante est d'obtenir $\mathbf{B}$ en ajustant seulement les covariances, donc les éléments non diagonaux de la variance, ce qui est raisonnable

pour les données réduites car la diagonale est fixée. Donc, quand $\mathbf{M} = \mathbf{I}_m$, on cherche numériquement le minimum $\min_{\mathrm{B}\, de\, rang\, 2} \sum_{j \neq g=1}^{m} (\mathbf{y}'_{(j)}\mathbf{W}\mathbf{y}_{(g)} - \mathbf{b}'_j\mathbf{b}_g)^2$ où $\mathbf{y}_{(j)}$ est la $j-^{me}$ colonne de $\mathbf{Y}$; puis on calcule les $\mathbf{a}_i$ par (10). Gabriel (1978b) a discuté cette approche sous le nom de « *Complex Correlational Biplot* » car sa solution peut inclure des composantes qui demandent une construction géométrique différente des produits scalaires (de telles composantes peuvent se produire car on n'approxime pas la matrice entière des variances $\mathbf{Y}'\mathbf{W}\mathbf{Y}$ mais seulement ses éléments non diagonaux). La « *Joint Correspondence Analysis* » de Greenacre (1988) utilise le même ajustement des covariances pour l'analyse des tableaux de contingence sans discuter la possibilité de solutions complexes. Evidemment, on pourrait y ajouter la représentation des individus par les $\mathbf{a}_i$ de (10).

3.10. Biplots par ajustement bilinéaire généralisé

On peut aussi généraliser les ajustements des biplots en utilisant les méthodes de régression bilinéaire généralisée où les pondérations et les métriques dépendent des espérances. Un tableau de contingence, par exemple, peut être ajusté et représenté en considérant ses fréquences comme des variables de Poisson, ce qui exige que les pondérations et les métriques soient des fonctions des espérances. Les calculs sont alors itératifs mais les principes et l'utilisation du biplot sont les mêmes que dans les exemples discutés ci-dessus (Falguerolles et Francis, 1992; van Eeuwijck, 1995; Gower et Hand, 1996, Section 11.3; Gabriel, 1998).

Il faut se rendre compte que l'ajustement en rang réduit n'est pas, en général, tout ce qu'on ajuste aux données. Pour l'*ACP* classique, par exemple, on commence en ajustant par les moyennes des variables, puis on ajuste les composantes principales aux résidus. Il serait élégant d'utiliser le même critère pour les deux ajustements (voir Gabriel, 1978a pour les moindres carrés simples et Gabriel, 1998 pour la régression généralisée). En pratique d'exploration cependant, il n'est pas toujours nécessaire de s'astreindre à cette élégance mathématique.

3.11. Biplots non linéaires

Gower et Hand (1996, chapitre 6) discutent les possibilités de représenter des données par indicateurs-lignes et indicateurs-colonnes qui échappent à la géométrie euclidienne aux axes rectilignes. On n'examine pas ici ces généralisations.

4. Le biplot comme diagnostic d'irrégularités et de régularités (structures)

4.1. L'exemple de l'orge au Minnesota

L'emploi du biplot comme outil diagnostique est illustré par l'exemple de récoltes d'orge au Minnesota (Immer, Hayes and Powers, 1934 ; voir les donnés dans Cleveland, 1993). Ces données de dix variétés d'orge plantées à six endroits en 1931 et en 1932 ont été l'objet de diverses analyses commençant avec celle de R. A. Fisher (1935).

La représentation en biplot exige que les données soient sous forme matricielle. Donc, s'il y a plus que deux facteurs il faut les combiner en deux groupes, l'un définissant les lignes de la matrice, l'autre ses colonnes. Ainsi les données d'orge ont été rangées dans une matrice de 12 lignes, une pour chaque combinaison (endroit,année), et 10 colonnes, une par variété.

Qualités : 0,8867 pour Y, 0,9514 pour Y'Y et YY'

FIG. 14. – Orges au Minnesota (données centrées sur la moyenne générale) : biplot SYM avec métrique euclidienne.

Le biplot de ces données, centrées sur la moyenne de la matrice entière (c'est le centrage préféré pour les diagnostics, voir Bradu et Gabriel, 1978), consiste

en 12 flèches qui sont les indicateurs-(endroit,année) et 10 points qui sont les indicateurs-variétés (Figure 14). Les coordonnées de ces indicateurs ont été calculées par la méthode des moindres carrés simples, ce qui produit une qualité de représentation de 88,67 %, dont 79,45 % sont expliqués par la variation horizontale du biplot et seulement 9,22 % par sa variation verticale. La mode de représentation de l'exemple dans cette section est celui des biplots SYM, ajustés avec des pondérations $\mathbf{W} = \mathbf{I}_{12}$ et la métrique $\mathbf{M} = \mathbf{I}_{10}$, dont les coordonnées sont $\mathbf{A}_{\{1/2\}} = [\mathbf{u}_{(1)}\mathrm{d}_1^{1/2} \quad \mathbf{u}_{(2)}\mathrm{d}_2^{1/2}]$ et $\mathbf{B}_{\{1/2\}} = [\mathbf{v}_{(1)}\mathrm{d}_1^{1/2} \quad \mathbf{v}_{(2)}\mathrm{d}_2^{1/2}]$. Les détails de ce mode de biplot seront discutés dans la Section 5 ; son interprétation est similaire à celle du biplot RMP et les diagnostics de structures sont les mêmes.

4.2. La découverte d'une irrégularité

On trouve (Figure 14) que tous les indicateurs-(endroit,année) sont dans une bande horizontale assez étroite et les indicateurs-variétés sont tous à droite de l'origine. Donc, les combinaisons (endroit,année) dont les indicateurs sont sur la droite du biplot ont eu en général des récoltes grandes, tandis que les indicateurs sur la gauche correspondent aux combinaisons (endroit,année) aux récoltes faibles. La combinaison des plus grandes récoltes fut $W1$ (Waseco, 1931) et celle des plus faibles récoltes $G2$ (Glabron, 1932).

On voit aussi que tous les indicateurs-variétés ont des coordonnées horizontales plus ou moins égales. Comme l'essentiel de la variation de ce biplot est horizontale, cela signifie que les différences entre les variétés ne sont pas très importantes. Néanmoins, si on examine les détails, on peut déceler trois groupes, *tre* et *wis* à une extrême, *sva*, *man*, *475*, *vel*, *gla*, *pea* à l'autre et *457*, *462* au milieu. Mais les différences entre ces groupes représentent seulement des petites différences de récolte.

Pour démêler les effets des endroits et des années on a ajouté un trait entre chaque indicateur de 1931 et celui du même endroit en 1932 (Figure 15). On aperçoit que ces segments ne sont pas d'égale longueur : ceux à droite sont en général plus longs que ceux à gauche. On conclut que la variabilité est plus grande pour les plus grandes récoltes, ce qui suggère d'effectuer une transformation logarithmique des données pour représenter la variation relative des récoltes.

Le biplot des logarithmes des récoltes, centrés sur la moyenne des logarithmes (Figure 16) a une qualité de représentation de 85,58 %, dont 76,91 % en direction horizontale et 8,67 % dans la direction verticale. La qualité et l'apparence de ce biplot sont très similaires à celles du biplot des récoltes elles-mêmes (Figure 14) si ce n'est que les différences de 1931 à 1932 changent moins d'une variété à l'autre.

Les segments 1931, 1932 sur le biplot des logarithmes (Figure 17) sont plus horizontaux que verticaux. Pour cinq endroits la direction 1931 à 1932 est de droite à gauche. Seul pour M (Morris) la direction est de gauche à droite, bien que ce segment soit à peu près de même longueur que les autres. Ainsi, on trouve que l'évolution proportionnelle des récoltes est plus ou moins la même

FIG. 15. – Orge au Minnesota (données centrées sur la moyenne générale) : biplot *SYM* avec métrique euclidienne et « intervalles » 1931-1932.

dans tous les endroits, à l'exception de Morris où son sens est inversé. C'est une découverte étonnante et difficile à accepter car les différences annuelles en récoltes s'expliquent en général par les variations du temps et cela ne peut avoir des directions inverses dans la même région. Il est plus plausible qu'il y a eu une erreur dans les données de Morris, les taux de 1931 ayant été échangés avec ceux de 1932. (Pour les détails sur ces données et l'explication de cette « anomalie Morris », voir le livre « *Visualizing Data* » de William Cleveland, 1993.) On remarque que la découverte de cette anomalie est très facile en regardant le biplot. Comme le dit Cleveland : « *Tools matter* ».

4.3. Les structures des régularités

En analyse de données, les biplots mènent souvent à la découverte d'erreurs, d'outliers ou autres irrégularités. On les découvre facilement car leurs indicateurs ne font pas partie des tendances générales de la plupart des indicateurs et quelquefois correspondent entièrement à une dimension. Le biplot initial d'une analyse, utilisant autant de dimensions que possible, indique souvent des irrégularités et l'étude propre des phénomènes observés ne peut être abordée qu'après résolution de ces irrégularités.

Mais le biplot ne mène pas seulement à la découverte d'irrégularités ; il est encore plus utile pour identifier des régularités qui peuvent indiquer des

Qualités : 0,8558 pour Y, 0,9484 pour Y'Y et YY'

FIG 16. – Orge au Minnesota (logarithmes centrés sur la moyenne générale) : biplot SYM avec métrique euclidienne.

structures. On a vu, par exemple, que les indicateurs-variétés des récoltes d'orge étaient près d'une droite. Cela permet de conclure (d'après Bradu et Gabriel, 1978; voir aussi Tableau 1) qu'un bon ajustement à ces données est possible par une structure de type Finlay-Wilkinson

$$y_{(endoit,année),variété} \approx \alpha_{(endroit,année)} + \gamma_{variété}\, \delta_{(endroit,année)}. \qquad (40)$$

De plus, on trouve que les indicateurs-(endroits,années) sont dans une bande étroite, ce qui permet (Tableau 1) de choisir une structure non additive à la Tukey (1949) qui peut être paramétrée (Mandel, 1961) comme

$$y_{(endroit,année),variété} \approx \eta + \gamma_{variété}\, \delta_{(endroit,année)}. \qquad (41)$$

Pour démêler les effets de l'endroit et de l'année, on a affecté un indicateur-moyenne à chaque endroit, le mettant à mi-chemin entre les indicateurs des deux années (Figure 18). Cette construction est justifiée parce que la moyenne de chaque endroit est une fonction linéaire des lignes de la matrice des données et peut être représentée par la même fonction linéaire des indicateurs-(endroits,années) (Gabriel, 1995a, Section 9.3.1). Pour les données d'orge, on trouve que ces indicateurs-endroits sont proches d'une droite qui est à peu

FIG. 17. – Orge au Minnesota (logarithmes centrés sur la moyenne générale) : biplot SYM avec métrique euclidienne avec intervalles 1931-1932.

près perpendiculaire à la droite dont les indicateurs-variétés sont proches, ce qui indique un bon ajustement par une structure additive :

$$\overline{y}_{endroit,variété} \approx \alpha_{endroit} + \gamma_{variété}. \tag{42}$$

Le fait que, sur ce biplot, les deux droites sont parallèles aux axes est sans importance car la représentation par biplot ne dépend pas des axes. Ainsi, les règles pour identifier des structures ne dépendent pas des axes, seulement des correspondances entre les indicateurs.

4.4. Quelques règles pour identifier les structures

Si on peut paramétrer la collection d'indicateurs-lignes et/ou la collection d'indicateurs-colonnes, on peut diagnostiquer une structure pour les données. La règle générale de modélisation par biplot est assez simple : si, pour chaque i, les indicateurs-lignes satisfont

$$\mathbf{a}_i' = \big(\phi_1(\pi_i), \phi_2(\pi_i)\big) \tag{43}$$

pour des fonctions $\phi_1(\cdot)$, $\phi_2(\cdot)$ et un paramètre π_i, et/ou si, pour chaque j, les indicateurs-colonnes satisfont

$$\mathbf{b}_j' = \big(\psi_1(\rho_j), \psi_2(\rho_j)\big) \tag{44}$$

FIG 18. – Orge au Minnesota (logarithmes centrés sur la moyenne générale) : biplot *SYM* avec métrique euclidienne moyennes des indicateurs de 1931 et 1932.

pour des fonctions $\psi_1(\cdot)$, $\psi_2(\cdot)$ et un paramètre π_i, il s'en suit que les données seront approximées par une structure

$$y_{ij} = \mathbf{a}_i'\mathbf{b}_j = \phi_1(\pi_i)\psi_1(\rho_j) + \phi_2(\pi_i)\psi_2(\rho_j). \tag{45}$$

Par exemple, si les **a**'s sont sur une droite horizontale et les **b**'s sur une droite verticale, cela peut être paramétré comme $\mathbf{a}_i' = (\pi_i, \pi_0)$ et $\mathbf{b}_j' = (\rho_0, \rho_j)$ et par conséquence on a la structure $y_{ij} = \mathbf{a}_i'\mathbf{b}_j = \rho_0\pi_i + \pi_0\rho_j$, c'est-à-dire une structure additive. De plus, on peut conclure la même chose si les **a**'s et les **b**'s sont sur deux droites perpendiculaires, quelle que soit leur relation aux axes. C'est parce que les caractéristiques du biplot sont invariantes par rotation et réflexion et donc que le diagnostic d'additivité reste le même si l'on tourne les droites pour les rendre parallèles aux axes (voir la discussion de (48) ci-dessous).

Certaines règles diagnostiques spéciales sont connues pour les biplots à deux dimensions (voir Bradu et Gabriel, 1978 et Tableau 1). D'autres (Tableaux 2 et 3) sont connues pour les biplots dans un espace à trois dimensions (Chuang, Gabriel et Therneau, 1986) et peuvent être visualisées sur l'écran avec des logiciels interactifs.

TABLEAU 1. – Diagnostics par biplots en deux dimensions (r_i et s_i sont des paramètres de lignes, u_j et v_j des paramètres de colonnes, M est une constante).

	$\mathbf{b}_j$'s colinéaires		Autres $\mathbf{b}_j$'s
Structure des $\mathbf{a}_i$	Perp. aux $\mathbf{a}_i$'s	Non perp. aux $\mathbf{a}_i$'s	
LINÉAIRE			
Affine	$r_i + v_j$	$M + s_i v_j$	$u_j + s_i v_j$
	Additif	*Tukey non additivité*	*Finlay-Wilkinson*
Vectoriel	v_j	$s_i v_j$	$s_i v_j$
		Rang 1	*Rang 1*
AUTRE		$r_i + s_i v_j$	$r_i u_j + s_i v_j$
	Sans structure	*Finlay Wilkinson*	*Rang 2*

(Bradu and Gabriel, 1978).

4.5. Des structures non linéaires

Les fonctions qu'on identifie sur un biplot et modélise en tant que structures ne sont pas nécessairement linéaires. Un exemple en est le biplot tridimensionnel de données de températures pendant 24 mois à 50 endroits étudiées par Tsianco et Gabriel (1984). Ces auteurs ont identifié une structure elliptique, de sorte que (44) devenait, pour le mois j,

$$\mathbf{b}'_j = \big(\mu, \alpha \cos(\rho_j),\ \beta \sin(\rho_j)\big), \tag{46}$$

ce qui mène, après un peu de manipulation, à une structure

$$y_{i,j} \longmapsto \eta_i + \xi_i + \psi_i \cos(\phi_i + \rho_j). \tag{47}$$

Ils ont pu identifier η_i et ψ_i avec la moyenne et l'amplitude de l'évolution annuelle de la température à l'endroit i, ξ_j comme l'effet du mois j, ϕ_i et ρ_j comme contributions de l'endroit et du mois à la phase (les estimateurs de ϕ_i étaient positifs pour l'hémisphère nord et négatifs pour l'hémisphère sud). C'était donc une structure assez logique pour ce phénomène.

5. Comment lever les indéterminations : les divers modes de biplots

Étant donné une métrique $\mathbf{M}$ et des pondérations $\mathbf{W}$, la méthode des moindres carrés approxime une matrice $\mathbf{Y}\boxed{(n \times m)}$ par une matrice unique $\mathbf{H}\boxed{(n \times m)}$ de rang deux. Quand on exprime cette approximation par une factorisation $\mathbf{AB}'$, les facteurs $\mathbf{A}\boxed{(n \times 2)}$ et $\mathbf{B}\boxed{(m \times 2)}$ ne sont pas uniques. En effet, à chaque matrice régulière $\mathbf{R}(2 \times 2)$ est associée une factorisation

$$\mathbf{AR}(\mathbf{BR}'^{-1})' = \mathbf{AB}' \tag{48}$$

TABLEAU. 2. – Diagnostics par biplots en 3-D de données en tableaux à deux dimensions (paramètres-lignes r = r_i, s = s_i, t = t_i ; paramètres-colonnes u = u_j, v = v_j, w = w_j, constante M).

	$\mathbf{b}_j$'s coplanaires		Autres $\mathbf{b}_j$'s
Structure des $\mathbf{a}_i$'s	Perp. aux $\mathbf{a}_i$'s	Non perp. aux $\mathbf{a}_i$	
PLANAIRE			
Affine	u + sv + t	M + sv + tw	u + sv + tw
Vectoriel	sv + t	sv + tw	sv + tw
LINÉAIRE			
Affine	u + t	M + tw	u + tw
Vectoriel	t	tw	tw
AUTRE			
	sans structure	r + sv + tw	ru + sv + tw

(Chuang, Gabriel and Therneau, 1986).

TABLEAU 3. – Diagnostics par biplots en 3-D de données en tableaux à trois dimensions (paramètres-lignes i et strates e r = $r_{i,e}$, s = $s_{i,e}$ et t = $t_{i,e}$; paramètres-colonnes j u = u_j, v = v_j et w = w_j ; paramètres-lignes R = R_i, S = S_i et T = T_i ; paramètres-strates R′ = R'_e, S′ = S'_e et T′ = T'_e ; constante M).

	$\mathbf{b}_j$'s coplanaires		Autres $\mathbf{b}_j$'s
Structure des $\mathbf{a}_{i,e}$'s	Perp. aux $\mathbf{a}_i$'s	Non perp. aux $\mathbf{a}_{i,e}$	
PLANS PARALLÈLES	Ru + sv + t	R + sv + tw	Ru + sv + tw
LIGNES PARALLÈLES	Ru + Sv + t	R + Sv + tw	Ru + Sv + tw
Plan affine	u + Sv + t	M + Sv + tw	u + Sv + tw
Plan vectoriel	Sv + t	Sv + tw	Sv + tw
STRUCTURES PARALLÈLES	sans structure	R + Sv + Tw +R′ + S′v + T′w	Ru + Sv + Tw +R′u + S′v + T′w
Plan affine	u + S + Tw + S′ + T′w	M + Sv + Tw + S′v + T′w	u + Sv + Tw + S′v + T′w
Plan vectoriel	S + Tw + s′ + T′w	Sv + Tw + S′v + T′w	Sv + Tw + S′v + T′w
TREILLIS			
Plan affine	u + Sv + T′	M + Sv + T′w	u + Sv + T′w
Plan vectoriel	Sv + (T + T′)w	Sv + T′w	Sv + T′w

différente, ce qui fournit des biplots différents de la même approximation, une représentation pour chaque **R**. Pour des matrices **R** orthogonales, le biplot

change par rotation et réflexion des **a**'s et des **b**'s. Pour d'autres **R**, il y aura d'autres changements, en particulier d'échelles, comme en (29) et (30) ci-dessus. On peut profiter de ces changements, qui n'ont pas d'effet sur les produits scalaires et l'approximation des données, pour choisir certains biplots qui sont spécialement intéressants parce qu'ils permettent aussi une bonne représentation de la variance multidimensionnelle $\mathbf{Y}'\mathbf{W}\mathbf{Y}$ ou de la forme $\mathbf{Y}\mathbf{M}\mathbf{Y}'$, comme on discute dans cette section.

La discussion suivante suppose des ajustements avec pondérations égales et métrique euclidienne, donc les matrices $\mathbf{W} = \mathbf{I}_n$ et $\mathbf{M} = \mathbf{I}_m$ seront omises. Bien entendu, ces conclusions peuvent être généralisées à d'autres **W** et **M**. Il faut insister sur le fait que c'est une discussion de qualité proportionnelle de représentation et non de qualité absolue, car le rôle exploratoire principal des graphiques est de comparer et non d'évaluer les valeurs absolues.

Dans la Section 2.2, ci-dessus, on a démontré, (13), (14), que les indicateurs-lignes d'un biplot peuvent être calculés comme lignes de

$$\mathbf{A}_{\{\lambda\}} \stackrel{def}{=} \left[\mathbf{u}_{(1)}\mathrm{d}_1^{\lambda} \quad \mathbf{u}_{(2)}\mathrm{d}_2^{\lambda}\right], \tag{49}$$

et les indicateurs-colonnes comme lignes de

$$\mathbf{B}_{\{\mu\}} \stackrel{def}{=} \left[\mathbf{v}_{(1)}\mathrm{d}_1^{\mu} \quad \mathbf{v}_{(2)}\mathrm{d}_2^{\mu}\right], \tag{50}$$

le choix des puissances λ et μ déterminant ce que nous appelons le mode (ou la modalité) du graphique. Les produits scalaires de ces indicateurs fournissent les approximations des éléments

$$y_{ij} \longmapsto \mathbf{a}'_{\{\lambda\}i}\mathbf{b}_{\{\mu\}j}, \tag{51}$$

ce qui correspond à (15). Les éléments de la forme sont approximés par

$$\mathbf{y}'_i\mathbf{y}_e \approx \mathbf{a}'_{\{\lambda\}i}\mathbf{a}_{\{\lambda\}e}, \tag{52}$$

correspondant à (19), donc les dissimilarités par

$$||\mathbf{y}_i - \mathbf{y}_e|| = \sqrt{\sum_{j=1}^{m}(y_{ij} - y_{ej})^2} \approx ||\mathbf{a}_{\{\lambda\}i} - \mathbf{a}_{\{\lambda\}e}||. \tag{53}$$

Pareillement (pour données centrées) les covariances sont approximées par

$$\mathbf{y}'_{(j)}\mathbf{y}_{(g)} = \sum_{i=1}^{n} y_{ij}y_{ig} \approx \mathbf{b}_{\{\mu\}j}\mathbf{b}'_{\{\mu\}g}, \tag{54}$$

correspondant à (24).

On a déjà remarqué que les approximations de ce type sont optimales, au sens des moindres carrés, dans les cas suivants :

si $\lambda + \mu = 1$ on obtient l'approximation optimale des données **Y**,

si $\lambda = 1$ on obtient l'approximation optimale de la forme $\mathbf{Y}\mathbf{Y}'$,

si $\mu = 1$ on obtient l'approximation optimale des covariances $\mathbf{Y}'\mathbf{Y}$.

Donc, cette méthode de construction ne peut pas produire un graphique qui présente simultanément les approximation optimales de $\mathbf{Y}$, de $\mathbf{YY}'$ et de $\mathbf{Y}'\mathbf{Y}$, mais peut construire des graphiques qui optimisent au maximum deux de ces approximations, comme suit.

$\underline{\lambda = 1}$ et $\underline{\mu = 0}$: le biplot RMP représente les approximations optimales $\mathbf{A}_{\{1\}}\mathbf{B}'_{\{0\}}$ de $\mathbf{Y}$ ainsi que $\mathbf{A}_{\{1\}}\mathbf{A}'_{\{1\}}$ de $\mathbf{YY}'$, mais son approximation $\mathbf{B}_{\{0\}}\mathbf{B}'_{\{0\}}$ n'est pas optimale pour $\mathbf{Y}'\mathbf{Y}$.

$\underline{\lambda = 0}$ et $\underline{\mu = 1}$: le biplot CMP représente les approximations optimales $\mathbf{A}_{\{0\}}\mathbf{B}'_{\{1\}}$ de $\mathbf{Y}$ et $\mathbf{B}_{\{1\}}\mathbf{B}'_{\{1\}}$ de $\mathbf{Y}'\mathbf{Y}$, mais son approximation $\mathbf{A}_{\{0\}}\mathbf{A}'_{\{0\}}$ n'est pas optimale pour $\mathbf{YY}'$.

$\underline{\lambda = \mu = 1/2}$: le biplot dit SYM (symétrique) est un compromis entre les biplots RMP et CMP. Son approximation $\mathbf{A}_{\{1/2\}}\mathbf{B}'_{\{1/2\}}$ de $\mathbf{Y}$ est optimale, mais ses approximations $\mathbf{A}_{\{1/2\}}\mathbf{A}'_{\{1/2\}}$ de $\mathbf{YY}'$ et $\mathbf{B}_{\{1/2\}}\mathbf{B}'_{\{1/2\}}$ de $\mathbf{Y}'\mathbf{Y}$ ne le sont pas.

$\underline{\lambda = \mu = 1}$: Le «graphique Benzécri» ou «représentation barycentrique» représente simultanément les approximations optimales $\mathbf{A}_{\{1\}}\mathbf{A}'_{\{1\}}$ de $\mathbf{YY}'$ et $\mathbf{B}_{\{1\}}\mathbf{B}'_{\{1\}}$ de $\mathbf{Y}'\mathbf{Y}$, mais son $\mathbf{A}_{\{1\}}\mathbf{B}'_{\{1\}}$ n'est pas optimale pour $\mathbf{Y}$. (Benzécri, 1973, a proposé ce graphique pour l'analyse de tableaux de fréquences, et Galindo Villardón, 1986, l'a généralisé pour des données quelconques en l'appelant biplot HJ).

$\underline{\lambda = \mu = 2/3}$: le «biplot 2/3» (Gabriel, 2002) représente $\mathbf{Y}$ par $\mathbf{A}_{\{2/3\}}\mathbf{B}'_{\{2/3\}}$, $\mathbf{YY}'$ par $\mathbf{A}_{\{2/3\}}\mathbf{A}'_{\{2/3\}}$ et $\mathbf{Y}'\mathbf{Y}$ par $\mathbf{B}_{\{2/3\}}\mathbf{B}'_{\{2/3\}}$. Aucune de ces approximations n'est optimale, mais toutes sont très proches de l'optimum, comme on va le montrer ci-dessous.

Il y a eu beaucoup de discussions sur les avantages et désavantages de chacun de ces graphiques (voir, par exemple, Greenacre, 1984, 1993 et Venables and Ripley, 1999, pp. 335-6). Les praticiens, d'autre part, ont remarqué que tous donnaient des conclusions similaires, bien que non identiques. La similarité était étonnante en tenant compte des différences de construction.

On peut illustrer cette similarité avec les données de récoltes d'orge. Chacun des graphiques de ces données, centrées sur leur moyenne générale (Figure 19) consiste en 12 points-(endroits,années) et 10 flèches-variétés. En comparant ces graphiques on remarque une différence très évidente entre les biplots RMP et CMP : le premier présente un nuage de points plus ou moins elliptique très aplati et un éventail très déployé de flèches ; le deuxième montre un nuage plus circulaire de points et un éventail plus serré de flèches ; le troisième graphique, le biplot SYM, est intermédiaire ; le quatrième, le graphique Benzécri, présente le nuage elliptique aplati du biplot RMP et l'éventail serré du biplot CMP. Le biplot 2/3 n'est pas montré car il ressemble fortement au biplot SYM.

Malgré leurs différences, ces graphiques ont beaucoup en commun. Chacun révèle le même ordre de flèches-variétés à partir de *sva* et *man* et jusqu'à *wis*

FIG 19. – Orge au Minnesota (données centrées sur la moyenne générale) : biplot divers avec métrique euclidienne.

et *tre*, et chacun montre les flèches avec abscisses positives. En ce qui concerne les indicateurs-(endroits,années), tous les graphiques ont le point $U1$ près de l'origine, c'est-à-dire qu'en 1931 les récoltes à University Farm étaient proches des récoltes moyennes. Pareillement, tous les graphiques montrent qu'il y a eu des petites récoltes de toutes les variétés à G (Grand Rapids) en 1932, et des grandes récoltes à W (Waseca) en 1931. Il existe encore d'autres similarités entre les quatre graphiques, mais il y a aussi des différences. Par exemple, le biplot RMP, qui a la meilleure qualité de représentation de la forme, suggère que $C1$ est plus similaire à $W2$ qu'à $U1$ tandis que le biplot CMP présente $C1$ comme plus proche de $U1$ que de $W2$. Une autre différence est que le biplot CMP, qui donne la meilleure représentation des variances et covariances, indique que les corrélations sont positives entre toutes les variétés, tandis que le biplot RMP suggère que certaines paires de variétés ont de faibles corrélations négatives.

Pour comparer les qualités proportionnelles des représentation de $\mathbf{Y}$, $\mathbf{YY}'$ et $\mathbf{Y}'\mathbf{Y}$ par les divers graphiques, on utilise la fonction

$$\psi(\Delta, v) = \frac{\{1 + \Delta^{1+v}\}^2}{\{1 + \Delta^2\}\{1 + \Delta^{2v}\}} \tag{55}$$

qui lie les qualités non optimales aux qualités optimales comme suit. Si d_1 et d_2 sont la première et la deuxième valeur singulière de $\mathbf{Y}$,

$$corr^2(\mathbf{A}_{\{\lambda\}}\mathbf{B}'_{\{\mu\}}, \mathbf{Y}) = corr^2\ (optimale\ pour\ \mathbf{Y})\ \psi\left(\frac{d_2}{d_1}, \lambda + \mu\right), \quad (56)$$

$$corr^2(\mathbf{A}_{\{\lambda\}}\mathbf{A}'_{\{\mu\}}, \mathbf{YY}') = corr^2\ (optimale\ pour\ \mathbf{YY}')\ \psi\left(\frac{d_2^2}{d_1^2}, \lambda\right), \quad (57)$$

et

$$corr^2(\mathbf{B}_{\{\mu\}}\mathbf{B}'_{\{\mu\}}, \mathbf{Y}'\mathbf{Y}) = corr^2\ (optimale\ pour\ \mathbf{Y}'\mathbf{Y})\ \psi\left(\frac{d_2^2}{d_1^2}, \mu\right), \quad (58)$$

(Gabriel, 2002). Evidemment, $\psi(\Delta, v)$ est une fonction de préservation de la qualité de représentation.

FIG. 20. – Contours de la fonction ψ de préservation de la qualité d'ajustement de $\mathbf{Y}$ (partie supérieure des contours) et de $\mathbf{YY}'$ ou $\mathbf{Y}'\mathbf{Y}$ (partie inférieure).

Des lignes de niveau de la fonction $\psi(\Delta, v)$ sont représentées sur la Figure 20. La partie supérieure de la figure concerne l'approximation de $\mathbf{Y}$ et ses paramètres sont $\Delta = d_2/d_1$ et $\nu = \lambda + \mu$; la partie inférieure concerne les approximations de $\mathbf{YY}'$ et de $\mathbf{Y}'\mathbf{Y}$ et ses paramètres sont $\Delta = d_2^2/d_1^2$ et $\nu = \mu$ ou $\nu = \lambda$. La figure donne aussi les qualités de certains modes d'approximation

comme, par exemple, la qualité de l'approximation de $\mathbf{Y}$ par le graphique Benzécri qui est représentée par une droite horizontale à la hauteur de $\nu = 2$.

La fonction $\psi(\Delta, v)$ atteint son maximum de $\psi = 1$ sur les deux côtés de la Figure 20, donc pour $\Delta = 0$ et pour $\Delta = 1$, soit pour $\mathrm{d}_2 = 0$ et pour $\mathrm{d}_2 = \mathrm{d}_1$. Si $d_2 = d_1$, c'est-à-dire si les deux premières dimensions sont de la même importance, alors tous les graphiques sont égaux et optimaux. La maximum $\psi = 1$ est atteint aussi le long d'une crête horizontale au niveau de $\nu = 1$. Donc, si $\nu = \lambda + \mu = 1$, c'est-à-dire si les approximations sont par moindres carrés, la qualité est maximale quelles que soient les valeurs singulières.

Quand ν s'éloigne de 1, le niveau de $\psi(\Delta, v)$ baisse : pour $\nu > 1$ il baisse lentement, avec un minimum à peu près à mi chemin entre $\Delta = 0$ et $\Delta = 1$; pour $\nu < 1$ il baisse plus rapidement, avec un minimum plus proche de $\Delta = 0$. Ce qui est d'importance pour l'utilisation des biplots est que, dans tout l'intervalle $1/2 \leqslant \nu \leqslant 2$ la surface ne descend jamais au dessous de 0,95 et dans l'intervalle $0 \leqslant \nu \leqslant 1/2$ elle n'est jamais au dessous de 0,50. Donc, pour tous les graphiques avec $\min(\lambda, \mu) \geqslant 0,5$, la préservation est toujours au moins 0,95, quel que soit $\mathrm{d}_2/\mathrm{d}_1$. Par contre, pour les graphiques avec $\lambda = 0$ (le biplot CMP) ou $\mu = 0$ (le biplot RMP), la préservation peut descendre vers 0,50, spécialement si d_2 est petit par rapport à d_1.

On apprend aussi que les approximations non optimales de $\mathbf{YY}'$ et de $\mathbf{Y}'\mathbf{Y}$ par le biplot SYM, de $\mathbf{Y}$ par le graphique Benzécri, et de tous trois par le biplot *2/3*, préservent presque toutes les qualités des approximations optimales, surtout le dernier qui préserve plus que 0,99 de la qualité de l'approximation optimale de $\mathbf{Y}$ et plus que 0,98 de $\mathbf{YY}'$ et de $\mathbf{Y}'\mathbf{Y}$.

Pour les données des récoltes d'orge on a utilisé des astérisques sur la Figure 20 pour indiquer les valeurs de préservation $\psi(\Delta, v)$. On voit que pour $\mathbf{Y}$ le graphique Benzécri préserve plus que 0,97, tandis que pour $\mathbf{YY}'$ et $\mathbf{Y}'\mathbf{Y}$ la préservation du biplot SYM est à peu près 0,97. Le biplot *2/3* préserve 0,993 de la qualité des approximations de $\mathbf{Y}$, $\mathbf{YY}'$ et $\mathbf{Y}'\mathbf{Y}$.

L'étude des lignes de niveau de la fonction de préservation suggère une construction additionnelle. On pourrait choisir λ et μ de manière que les préservations des trois objectifs soient aussi proches et grandes que possible simultanément, c'est-à-dire qu'on chercherait les valeurs λ et μ qui donnent le maximum

$$\max_{\lambda,\mu} \left[\min_{\delta} \left\{ \psi(\delta, \lambda + \mu), \psi(\delta^2, \lambda), \psi(\delta^2, \mu) \right\} \right]. \tag{59}$$

On appellera le graphique construit avec la solution $\breve{\lambda} = \breve{\mu}$ de (53) le biplot *Maximin*; on trouve que les lieux de ses préservations pour $\mathbf{Y}$ et pour $\mathbf{YY}'$ et $\mathbf{Y}'\mathbf{Y}$ sont des courbes liées par l'égalité

$$\psi(\delta, 2\breve{\lambda}) = \psi(\delta^2, \breve{\lambda}) \tag{60}$$

(voir Figure 20). On constate que le biplot *Maximin* préserve toujours presque 0,99 des qualités optimales pour toutes les trois approximations, et que, à part

des cas de valeurs très basses de d_2/d_1, les préservations du biplot *Maximin* des données d'orge avec $\check{\lambda} = 0,6641$ sont très proches à celles du biplot *2/3*.

On peut conclure que le graphique Benzécri et les biplots *SYM*, *2/3* et *Maximin* ont des qualités de représentation proches de l'optimalité même pour les objectifs qu'ils ne visent pas optimalement. Les biplots *RMP* et *CMP*, d'autre part, peuvent perdre jusqu'à 50 % de la qualité de représentation optimale pour certains objectifs. Il ne faudra se servir de ces biplots que si le taux d_2/d_1 n'est pas trop faible ou si un objectif est d'intérêt particulier (comme, par exemple, les lignes en *MANOVA* qui représentent les échantillons à comparer) et l'autre objectif n'est que d'intérêt secondaire.

En ce qui concerne le choix entre le graphique Benzécri et les biplots *SYM*, *2/3* et *Maximin*, leurs différences ne sont pas importantes. Si l'on devait choisir, on devrait préférer le dernier, ou bien le biplot *2/3* qui est presque égal au dernier mais est plus facile à construire. Le biplot *SYM* a un petit avantage pour la représentation des données mêmes, le graphique Benzécri un petit avantage pour la représentation de la variance et de la forme, c'est-à-dire des distances, tandis que les biplots *2/3* et *Maximin* sont des compromis et sont presque égaux. Donc, la rivalité entre les défenseurs des biplots et les défenseurs du graphique de Benzécri n'est pas d'importance pratique.

La littérature est pleine de propositions de méthodes d'analyse et de représentation des données, chaque auteur prônant sa favorite, souvent en montrant que ses résultats ressemblent à ceux de méthodes bien connues. Il apparaît que le message des données surgit fréquemment sans être trop influencé par la méthodologie : *vincit omnia veritas.*

6. La qualité de représentation et le choix de dimension

Ayant vu ce que sont les biplots et comment les construire, posons-nous quelques questions sur l'objet de la représentation et sur le choix du nombre de dimensions qui sert à bien révéler ce qu'on cherche.

En statistique l'objet de la représentation graphique dépend du caractère des données. Pour les observations multidimensionnelles $y_{i,j}$, $i = 1, \ldots, n$; $j = 1, \ldots, m$, il faut distinguer d'un côté les observations sur des collections d'*individus identifiés*, $i = 1, \ldots, n$, et de l'autre côté les *échantillons aléatoires* de n individus provenant d'une distribution. En général, la distinction est la même que celle entre une matrice de données $\mathbf{Y}$ aux lignes étiquetées et aux lignes sans labels ou aux labels sans signification informative.

Deux exemples pourraient clarifier ces distinctions selon l'origine des données.

Un est la consommation de certaines protéines dans les pays européens (voir Fig. 8.3 de Gabriel, 1981) ce qui est une collection de soi-disant « individus » bien identifiés qu'il serait peu plausible de regarder comme un échantillon aléatoire. L'autre exemple est la collection des données de Lubischew (1962) sur 74 insectes parvenant de sources diverses. Pour l'analyse, on décide de la regarder comme un échantillon aléatoire et y appliquer les tests de signification (Caussinus, Hakam et Ruiz-Gazen, 2002, 2003).

La distinction entre données de collections identifiées et celles d'échantillons aléatoires concerne non seulement le choix de graphiques mais aussi la logique d'inférence statistique. Pour les collections identifiées les seules erreurs sont celles de mesure et d'observation et il est approprié de présumer un modèle à effets fixes (« fonctionnel » selon Fine, 1992),

$$y_{i,j} = \eta_{i,j} + e_{i,j}, \tag{61}$$

où les e's sont indépendants, d'espérances 0 et de variances σ^2 (inconnue). Sous cette hypothèse, l'intérêt principal devrait être de connaître les valeurs $\eta_{i,j}$ ou au moins connaître leurs régularités et irrégularités. Par exemple, dans le cas des chiens et des loups on s'intéressait aux diverses races des animaux et essayait de décrire leurs relations aux variables mesurées. Dans l'exemple des récoltes d'orge on se demandait s'il y avait une structure systématique (et on apercevait incidemment l'anomalie Morris), puis si elle pouvait être bien décrite par une structure additive $\eta_{i,j} = \alpha_i + \beta_j$, ou par une autre structure. Supposer un échantillonnage aléatoire aurait été complètement inopportun pour cet exemple. Avec de telles données on essaye de connaître les caractéristiques des $\eta_{i,j}$ sans être leurré pas les e's ; le mode RMP convient donc mieux que le mode CMP, si on doit choisir entre ces deux.

Pour les échantillons aléatoires, d'autre part, il faut supposer un modèle aléatoire (« structurel » selon Fine, 1992) où les vecteurs $\mathbf{y}_i' = (y_{i,1}, \ldots, y_{i,j}, \ldots, y_{i,m})$ sont n réalisations indépendantes provenant d'une distribution (qui peut être un mélange inconnu de lois de nature donnée). Sous de telles hypothèses le seul intérêt est de connaître les caractéristiques de la distribution. Les variances et les corrélations sont intéressantes comme paramètres de la distribution et, par conséquent, le mode CMP a un avantage sur le mode RMP (voir, par exemple, Corsten et Gabriel, 1973, qui analysent les données de pluie en m endroits pour n journées, et s'intéressent exclusivement à la distribution m-variable des pluies et non aux journées individuelles.) Par contre, sous l'hypothèse alternative d'une collection identifiée d'individus aux observations perturbées d'erreur, les variances et les corrélations ne sont que des sommaires descriptifs sans intérêt intrinsèque.

Pour les données d'origine aléatoire, les graphiques devraient servir à élucider le caractère de la distribution ou du mélange de distributions. Donc le rang est une caractéristique importante et les tests de dimension ont un rôle à jouer dans le choix des dimensions nécessaires pour représenter la distribution. S'il y a deux ou trois dimensions significatives un biplot doit avoir deux ou trois dimensions pour une description complète. S'il y en a davantage, un biplot ne donnera qu'une description partielle.

Pour les données d'individus identifiés et donc pour le modèle à effets fixes, les graphiques devraient approximer les espérances, non pas les observations elles-mêmes. La représentation des erreurs n'a aucun intérêt. De plus, la vraie dimension n'est pas d'importance : une ligne ou un plan, ou bien un espace tridimensionnel, peuvent donner de meilleures approximations que l'espace de toutes les « vraies dimensions ». Il est même possible que l'information représentée sur le biplot soit meilleure que celle des éléments des données

elles-mêmes qui sont plus influencés par les erreurs (voir, par exemple, Gauch, 1992). La question de signification des dimensions ne se pose donc pas avec ce type de données. Au lieu des tests on devrait avoir recours à la validation croisée qui mesure la qualité d'un biplot au moyen des prédictions qu'il fournit.

Pour la pratique, avec certains types de données aléatoires, on trouve des tests de signification de la dimension dans les travaux récents de Caussinus, Hakam et Ruiz-Gazen (2002, 2003). D'autre part, pour des données identifiées, on a besoin d'une bonne méthode de validation croisée des approximations en rang réduit (Jolliffe, 2002, 6.1.5). Les méthodes de ce genre sont plus compliquées que celles de la validation croisée d'une statistique unidimensionnelle, parce que les approximations de rang réduit sont définies pour des matrices complètes, et leurs redéfinition pour des matrices où manque un élément sont ambiguës. Des tentatives de définitions pour la validation croisée de l'ACP, donc pour les approximations en rang réduit, ont été faites par Wold (1976) et par Eastment et Krzanowski (1982), et nous en suggérons une autre (voir l'Appendice). Nous espérons publier bientôt une étude comparative de ces diverses méthodes de validation croisée (Gabriel, 2003b).

Le choix de rang d'approximation par validation croisée peut être illustré par les données de récoltes d'orge. On trouve (Tableau 4) que l'ajustement en rang deux est le meilleur, mais celui de rang un est presque aussi bon. La deuxième dimension des biplots de ces données ne peut ajouter que très peu. En effet, ce qu'on a découvert ci-dessus sur l'anomalie des données d'orge de Morris dépend entièrement de la dimension horizontale des biplots; la dimension verticale n'y contribue en rien. On aurait aussi bien vu ce phénomène sur un graphique unidimensionnel.

TABLEAU 4. – Coefficients de qualité d'ajustement pour des rangs divers calculés par validation croisée : données de récoltes d'orge.

	Rang r					
	1	2	3	4	5	6
Corrélation carrée	0,8332	0,8696	0,8545	0,8506	0,8338	0,8246

On trouve fréquemment des données qui sont mieux représentées en peu de dimensions qu'en beaucoup. Ainsi, deux dimensions sont souvent le meilleur choix ou très proche de celui-ci. Cela pourrait être dû à l'hétérogénéité de beaucoup de collections de données; par contre, il est possible que la représentation de données d'échantillons d'une seule population exige davantage de dimensions.

Quand on veut profiter de plus de dimensions, on peut avoir recours aux biplots multidimensionnels dont les calculs sont analogues à ceux en deux dimensions mais la visualisation est plus compliquée. Elle peut utiliser des rotations sur l'écran de l'ordinateur ou bien la collection de projections bidimensionelles. Par exemple, le biplot tridimensionnel d'une petite matrice de 9 lignes et 4 colonnes est montré dans la Figure 21 qui présente toutes les projections planes du biplot tridimensionnel. On voit bien l'analogie avec

les graphiques de tous les nuages plans comme celui de la Figure 5 ci-dessus. Néanmoins, dans le cas présent, les axes correspondent aux axes principaux, tandis que les variables sont représentées par des flèches. On remarque l'outlier '2' qu'on n'aurait pas vu sur le biplot bidimensionnel (premier panneau de la Figure 21).

L'avantage du biplot, par contraste avec la liste des coefficients de l'ACP, est qu'il montre l'ensemble sur le plan, ou l'espace, et pas seulement les grandeurs individuelles le long des axes. De la même manière, le biplot multidimensionnel permet une visualisation globale. Donc, quand on se sert du graphique des projection planes, il faudra soigneusement regarder les projections sur tous les plans principaux, pas seulement sur les axes 1 et 2, 3 et 4, etc.

FIG. 21. – Biplot tridimensionnel, représentation par projections bidimensionnelles. Un exemple.

7. Problèmes à résoudre

7.1. Transformations, liens d'espérances et fonctions de variance

En analyse confirmatoire univariable on choisit des transformations ou, alternativement, des fonctions de lien pour l'espérance et des fonctions de variance, afin de satisfaire des hypothèses d'un modèle probabiliste de la variabilité et d'une structure des espérances. Selon les tests d'ajustement et les graphiques des résidus on décide si les hypothèses sont appropriées et on les change si cela paraît nécessaire. L'application d'une pareille approche à la représentation graphique de données multidimensionnelles est plus ambiguë parce qu'elle sert à l'exploration de situations où l'on manque d'hypothèses claires sur la structure et où le modèle probabiliste est douteux.

Par exemple, une analyse de données de taches sur des feuilles de plantes a utilisé un certain modèle bilinéaire généralisé (Gabriel, 1998) ; mais était-ce le meilleur modèle ? Ou bien est-ce que la transformation logarithmique des données d'orge discutées plus haut (Section 4.1) est la meilleure pour visualiser les caractéristiques de ces données ? Un exemple de représentation trompeuse est celui où les données proviennent d'un modèle d'association RC (Goodman, 1991) et où on utilise une représentation sans transformation logarithmique, ce qui donne l'impression de trop de dimensions (effet Guttman). Le contraire est valable pour des données venant d'un modèle de corrélation (Baccini, Caussinus et de Falguerolles, 1993) où la représentation logarithmique est trompeuse. Comment choisir quand l'objet de la représentation est d'indiquer le modèle sans le connaître *a priori* ? Même les critères de qualité d'ajustement sont problématiques car l'objet de l'exploration n'est pas unique : on veut explorer des aspects différents des données comme outliers, clusters, structures bilinéaires, etc. Comment définir un abord optimal quand l'idée même de l'exploration est d'avoir plusieurs objectifs *a priori* inconnus ?

Donc on ne sait que procéder par tâtonnement.

7.2. La représentation de plusieurs matrices

Le problème de la comparaison de matrices similaires, comme celles de répétitions à des temps successifs et/ou à des endroits différents, a été abordé par plusieurs chercheurs. D'aucuns se sont concentrés sur des modèles tridimensionnels (voir, par exemple, Carlier et Kroonenberg, 1996), tandis que d'autres ont étudié directement la pluralité des graphiques (par exemple, Gheva, 1986). C'est un sujet qu'il faut approfondir et qui pourrait être utile pour l'étude exploratoire de grandes quantités de données comme en data mining.

7.3. Qualités d'ajustements proportionnels et absolus

Pour les données, leurs dissimilarités et leurs variances, ce sont effectivement les comparaisons qui sont de premier intérêt. Mais, pour les corrélations, les valeurs elles-mêmes sont souvent d'intérêt. (Par exemple, on ne voudra pas

seulement savoir si la corrélation entre X et Y est plus grande que celle entre U et W mais aussi si la première est positive et la dernière négative : sur le biplot, on verra si les angles entre leurs flèches sont, respectivement, aigus et obtus). Dans ces cas, ce qui est important c'est la qualité de l'ajustement absolu, pas de l'ajustement proportionnel. Donc, c'est à l'avantage du biplot *CMP*.

8. La définition du biplot et les conclusions

Ayant discuté les caractéristiques de ces graphiques, on pourrait revenir sur le problème de leur définition. Evidemment un biplot est une représentation graphique d'une matrice consistant en indicateurs-lignes et indicateurs-colonnes. Leurs produits scalaires étaient l'essentiel de la définition originale du biplot (Gabriel, 1971) comme représentation d'une approximation *optimale* des éléments de la matrice.

On a vu que les biplots *RMP*, *CMP* et *SYM* satisfont la définition de l'approximation optimale des données par produits scalaires ; on a vu aussi que le graphique Benzécri et les biplots *2/3* et *Maximin* ne satisfont pas l'optimum, mais que leurs produits scalaires donnent de très bonnes approximations. En pratique il faudra donc les inclure dans la définition et nous proposons donc une nouvelle définition du biplot comme graphique de $\mathbf{Y}$ au moyen d'indicateurs-lignes et indicateurs-colonnes dont les produits scalaires sont des *bonnes* approximations des éléments de $\mathbf{Y}$. Cette définition inclut non seulement la représentation d'ajustements par moindres carrés mais aussi par autres critères, comme ceux des ajustements robustes.

On remarque que cette définition est basée entièrement sur le graphique comme approximation des données et non sur une géométrie particulière comme celle des projections sur sous-espaces. Elle ne modélise pas les données pour visualiser leurs deux premières composantes comme le fait par exemple de Falguerolles (1998), mais on se concentre sur l'approximation des données elles-mêmes sur un plan.

Notre abord est principalement descriptif. C'est qu'à notre avis la visualisation doit précéder l'analyse, ou bien alterner avec elle, d'où notre présentation du biplot non comme méthode analytique mais comme outil visuel. Si nous préconisons le biplot comme outil de présentation c'est à cause de l'efficacité de sa visualisation. Une illustration en est la facilité de la découverte de l'anomalie Morris par les biplots (Section 4, ci-dessus) en comparaison avec les analyses géniales mais compliquées par Cleveland (1993). Nous sommes d'accord avec lui que « tools matter » et nous préconisons le biplot parce qu'il est facile à visualiser et comprendre et qu'il rend à l'œil son rôle important en analyse et interprétation.

Remerciements

Mes remerciements vont à Henri Caussinus et Dan Bradu qui m'ont aidé à mieux comprendre les problèmes discutés ici.

RÉFÉRENCES

ALLEN D. M. (1974). "The relationship between variable selection and data augmentation as a method of prediction." *Technometrics*, **16**, 125-127.

BACCINI A., CAUSSINUS H. et FALGUEROLLES A. de (1993). "Analysing dependence in large contingency tables : Dimensionality and patterns in scatter-plots." *Multivariate Analysis : Future Directions 2* (eds. C.M. Cuadras et C.R. Rao). Amsterdam; Elsevier, pp. 245-263.

BENZÉCRI J.-P. (1973). *L'Analyse des Données, Tome 2 : L'Analyse des Correspondances.* Paris : Dunod.

BENZÉCRI J.-P. (1978). «Sur l'analyse des tableaux binaires associés à une correspondance multiple.» *Cahiers de l'Analyse des Données*, **2**, 55-71.

BRADU D. et GABRIEL K. R. (1978). "The biplot as a diagnostic tool for models in two-way tables." *Technometrics*, **20**, 47-68.

BRUTSCHY B. et LEBART L. (1991). "Contiguity analysis and projection pursuit." *Applied Stochastic Models and Data Analysis* (eds. R. Gutierrez et M.J.M.Valderrama). World Scientific, pp.117-128.

CARLIER A. et KROONENBERG P.M. (1996). "Decompositions and biplots in three-way correspondence analysis." *Psychometrika*, 61, 355-373.

CAUSSINUS H. (1986). "Models and uses of principal component analysis.," en *Multidimensional Data Analysis*, (eds. J. de Leeuw, W. Heiser, J. Meulman et F. Critchley). Leiden, DSWO Press, pp. 149-170.

CAUSSINUS H. (1992). «Projections révélatrices», en *Modèles pour l'Analyse des Données Multidimensionnelles*, (eds. J. J. Droesbeke, B. Fichet et P. Tassi). Paris, Economica, pp. 241-265.

CAUSSINUS H. (1993). «Modèles probabilistes et analyse des données multidimensionnelles». *Journal de la Société de Statistique de Paris*, **134**, 15-32.

CAUSSINUS H. et FALGUEROLLES A. de (1987). «Tableaux carrés : modélisation et méthodes factorielles» *Revue de Statistique Appliquée*, **XXXV**, 35-52.

CAUSSINUS H. et FERRÉ L. (1992). "Comparing the parameters of a model for several units by means of principal components analysis." *Computational Statistics and Data Analysis*, 13, 269-280.

CAUSSINUS H., HAKAM S., et RUIZ-GAZEN A. (2002). «Projections révélatrices contrôlées – recherche d'individus atypiques.» *Revue de Statistique Appliquée*, **L**(4), 81-94.

CAUSSINUS H., HAKAM S., et RUIZ-GAZEN A. (2003). «Projections révélatrices contrôlées – groupements et structures complexes.» *Revue de Statistique Appliquée*, LI(1), 37-58.

CAUSSINUS, H. et RUIZ A. (1990). "Interesting projections of multidimensional data by means of generalized principal component analysis." *COMPSTAT 90.* Heidelberg : Physica Verlag, pp. 121-36.

CAUSSINUS H. et RUIZ-GAZEN A. (1993). "Projection pursuit and generalized principal component analysis," in *New Directions in Statistical Data Analysis and Robustness.* Basel : Birkhausen Verlag, pp.35-46.

CAUSSINUS H. et RUIZ-GAZEN A. (1995). "Metrics for finding typical structures by means of principal component analysis," in *Data Science and Its Applications.* Harcourt Brace, pp. 177-92.

CHUANG J.L.C., GABRIEL K. R., et THERNEAU T. M. (1986). *Use of 3-D biplots for diagnosing models to fit high-dimensional data.* University of Rochester, Department of Statistics Technical report 86/06.

CLEVELAND W. S. (1993). *Visualizing Data.* Murray Hill, NJ, AT&T Bell.

CORSTEN L.C.A. et GABRIEL K. R. (1976). "Graphical exploration in comparing variance matrices." *Biometrics*, **32**, 851-863.

DAIGLE G. et RIVEST L. P.(1992). "A robust biplot." *Canadian Journal of Statistics*, 20, 241-255.

EASTMENT H.T. et KRZANOWSKI W.J. (1982). "Cross-validatory choice of the number of components from a principal component analysis". *Technometrics*, **24**, 73-77.

ESCOUFIER Y. (1973). «Le traitement des variables vectorielles». *Biometrics*, **29**, 751-760.

FALGUEROLLES A. de (2000). «Trop de camemberts tuent le camembert». *Journal de la Société Française de Statistique*, **141**, 45-49.

FALGUEROLLES A. de (1998). "Log-bilinear biplots in action," in *Visualization Of Categorical Data* (eds. J. Blasius and M. Greenacre). San Diego : Academic Press, 527-539.

FALGUEROLLES A. de et FRANCIS B. (1992). "Algorithmic approaches for fitting bilinear models," in *COMPSTAT 92 : Computational Statistics*, Vol. 1, (eds. Y. Dodge and J. Whittaker). Heidelberg, Physica-Verlag, pp. 77-82.

FARAJ A. (1993). «Analyse de contiguïté : une analyse discriminante généralisée à plusieurs variables qualitatives». *Revue de Statistique Appliquée*, **41**, 73-84.

FARAJ A. (1994). "Interpretation tools for generalized discriminant analysis," in *New Approaches in Classification and Data Analysis* (eds. E, Diday, Y. Lechevallier, M. Schader, P. Bertrand et B. Burtschy). Berlin, Springer-Verlag, 285-291.

FARAJ A., et CAILLY F. (2001). "Spatial contiguity analysis : a method for describing spatial structures of seismic data." *Journal of Petroleum Science and Engineering*, **31**, 93-111.

FERRÉ L. (1995). "Improvement of some multidimensional estimates by reduction of dimensionality." *Journal of Multivariate Analysis*, **54**, 147-162.

FINE J. (1992). «Modèles fonctionels et structurels», en *Modèles pour l'Analyse des Données Multidimensionelles*, (eds. J. J. Droesbeke, B. Fichet et P. Tassi). Paris, Economica, pp. 21-60.

FINLAY K.W. et WILKINSON G.N. (1963). "The analysis of adaptation in a plant breeding programme," *Australian Journal of Agricultural Research*, **14**, 742-754.

FISHER R. A. (1935). *The Design of Experiments.* Edinburgh, Oliver and Boyd.

FRIENDLY M., et DENIS D. (2000). "The roots and branches of modern statistical graphics." *Journal de la Société Française de Statistique*, **141**, 51-60.

GABRIEL K. R. (1971). "The biplot – graphical display of matrices with application to principal component analysis." *Biometrika*, **58**, 453-467.

GABRIEL K. R. (1972). "Analysis of meteorological data by means of canonical decomposition and biplots." *Journal of Applied Meteorology*, **11**, 1071-77

GABRIEL K. R. (1978a). "Least squares approximation of matrices by additive and multiplicative models." *Journal of the Royal Statistical Society, B*, **40**, 186-196.

GABRIEL K. R. (1978b). "The complex correlational biplot," en *Theory Construction and Data Analysis in the Behavioral Sciences* (ed. S. Shye). San Francisco, Jossey-Bass, pp. 350-370.

GABRIEL K. R. (1981). "Biplot display of multivariate matrices for inspection of data and diagnosis," en *Interpreting Multivariate Data* (ed. V. Barnett). Chichester, Wiley, pp. 147-173.

GABRIEL K. R. (1982). "Biplot," en *Encyclopedia of Statistical Sciences*, Vol. I (eds. S. Kotz and N.L. Johnson). New York : Wiley, pp. 262-265.

GABRIEL K. R. (1987). "Some thoughts on comparing multivariate data with the map locations at which they were observed," en *Proceedings of the 19th Symposium on the Interface : Computer Science and Statistics* (eds. R. M. Heiberger et M. T. Martin). Alexandria, Virginia : American Statistical Association, pp. 139-146.

GABRIEL K. R. (1988). "Relating multivariate data to geographical location," en *Data Analysis, Expert Knowledge and Decisions* (eds. W. Gaul et M. Schader). Berlin : Springer, pp. 341-354.

GABRIEL K. R. (1995a). "Biplot display of multivariate categorical data, with comments on multiple correspondence analysis," en *Recent Advances in Descriptive Multivariate Analysis* (ed. W.J. Krzanowski). Oxford : Clarendon. Chapter 9.

GABRIEL K.R. (1995b). "MANOVA biplots for two-way contingency tables," in *Recent Advances in Descriptive Multivariate Analysis* (ed. W.J. Krzanowski). Oxford : Clarendon. Chapter 10.

GABRIEL K.R. (1997). "The effect of metrics on biplot display," en *Statistical and Mathematical Modelling in the Fields of Food Science, Biotechnology and Environment*, (ed. Ute Rmisch). Berlin : Frauenhofer Institut.

GABRIEL K.R. (1998). "Generalized bilinear regression." *Biometrika*, **85**, 689-700.

GABRIEL K.R. (2002). "Goodness of fit of biplots and correspondence analysis." *Biometrika*, **89**, 423-436.

GABRIEL K.R. (2003a). "Lower rank fits and biplots from incomplete data." (In preparation).

GABRIEL K.R. (2003b). "Cross-validation and choice of dimension for lower rank fits and biplots." (In preparation).

GABRIEL K.R., GALINDO M.P. et VICENTE-VILLARDÓN, J.L. (1998). "Use of Biplots to diagnose independence models in contingency tables," in *Visualization Of Categorical Data* (eds. J. Blasius and M. Greenacre). San Diego : Academic Press, 391-404.

GABRIEL K.R. et ODOROFF C. L. (1984). "Resistant lower rank approximation of matrices," in *Data Analysis and Informatics III*, (eds. E. Diday, M. Jambu, L. Lebart, J. Pages, and R. Tomassone). Amsterdam : North-Holland, 23-30.

GABRIEL K.R. et ODOROFF C. L. (1990). "Biplots in biomedical research." *Statistics in Medicine*, **9**, 469-485.

GABRIEL K.R. et ZAMIR S. (1979). "Lower rank approximation of matrices by least squares with any choice of weights." *Technometrics*, **21**, 489-498.

GALINDO VILLARDÓN P. (1986). "Una alternativa de representacion simultanea : HJ-biplot." *Questio*, **10**, 13-23.

GAUCH H. G. (1992). *Statistical Analysis of Regional Yield Trials : AMMI Analysis of Factorial Designs.* Amsterdam : Elsevier.

GAUCH H. G. et ZOBEL R. W. (1997). "Identifying mega-environment and targeting genotypes." *Crop Science*, **37**, 311-326.

GEARY R. C. (1954). "The contiguity ratio and statistical mapping." *The Incorporated Statistician*, **5**, 115-154.

GHEVA D. (1986). "The biplot graphic technique for the representation of multivariate time series." *European Journal of Operational Research*, **27**, 95-103.

GOODALL C. (1991). "Procrustes methods in the analysis of shape." *Journal of the Royal Statistical Society, Series B*, **53**, 285-339.

GOODMAN L. A. (1991). "Measures, models, and graphical displays in the analysis of cross-classified data." *Journal of the American Statistical Association*, **86**, 1085-1138.

GOWER J.C. (1971). "Statistical methods of comparing different multivariate analyses of the same data," in *Mathematics in the Archeological and Historical Sciences* (eds. F.R. Hodson, D.G. Kendall, et P. Tautu). Edinburgh : University Press, 138-149.

GOWER J. C. et HAND D. J. (1996). *Biplots.* London : Chapman and Hall.

GREENACRE M. J. (1984). *Theory and Application of Correspondence Analysis.* London : Academic Press.

GREENACRE M. J. (1988). "Correspondence analysis of multivariate categorical data by weighted least squares." *Biometrika*, **75**, 457-467.

GREENACRE M. J. (1993). "Biplots in Correspondence Analysis." *Journal of Applied Statistics*, **20**, 251-269.

HABER M. (1975). *The Singular Value Decomposition of Random Matrices.* Thèse de Ph.D., Jerusalem, Hebrew University.

HENRION R. and HENRION G. (1995). *Multivariate Datenanalyse.* Berlin, Springer.

HEO M. (1996). *On the fit of sample graphical displays to patterns in populations.* University of Rochester, Ph.D thesis.

HEO M. et GABRIEL K. R. (2001). "The fit of graphical displays to patterns of expectations." *Computational Statistics and Data Analysis*, **36**, 47-67.

HOUSEHOLDER A.S. et YOUNG G. (1938). "Matrix approximation and latent roots." *American Mathematical Monthly*, **45**, 165-171.

IMMER F. R., HAYES H. K. et LE ROY POWERS (1934). "Statistical determination of barley varietal adaptation." *Journal of the American Society of Agronomy*, **26**, 403-419.

JAMBU M. (1977). «Sur l'utilisation conjointe d'une classification hiérarchique et de l'analyse factorielle en composantes principales». *Revue de Statistique Appliquée*, **25**, 4, 5-35.

JOLLIFFE I. T. (2002). *Principal Component Analysis, 2e ed.* Springer, New York.

KEMPTON R. A. (1984). "The use of biplots in interpreting variety by environment interactions." *Journal of Agricultural Science*, **103**, 123-135.

LEBART L. (1969). *Analyse Statistique de la Contiguïté*, Publ. de l'ISUP, XVIII, 81-112.

LEBART L. (2001). «Classification, analyse de contiguïté et plus proches voisins», *Journées de Statistique XXVII, Nantes*, Paris : ASU, 503-506.

LINGOES J. C. et SCHÖNEMANN P. H. (1974). "Alternative measures for fit for the Schnemann-Carroll matrix fitting algorithm." *Psychometrika*, **39**, 423-427.

LUBISCHEW A. A. (1962). "On the use of discriminant functions in taxonomy." *Biometrics*, **18**, 455-477.

MANDEL J. (1961). "Non-additivity in two-way analysis of variance." *Journal of the American Statistical Association*, **56**, 878-888.

RAMSAY J. O., ten BERGE, J., et STYAN G. P. H. (1984). "Matrix correlation." *Psychometrika*, **49**, 403-423.

RUIZ-GAZEN A. (1996). "A very simple robust estimator for a dispersion matrix." *Computational Statistics and Data Analysis*, **21**, 463-473.

SAPORTA G. (1990). *Probabilités, Analyse des Données et Statistique.* Paris, Dunod.

SPARKS R., ADOLPHSON A. et PHATAK A. (1997). "Multivariate process monitoring using the dynamic biplot." *International Statistical Review*, **65**, 325-349.

ter BRAAK C.J.F.(1990). "Interpreting canonical correlation analysis through biplots of structure and weights." *Psychometrika*, **55**, 519-532.

ter BRAAK C.J.F.(1994). "Biplots in reduced rank regression." *Biometrical Journal*, **8**, 983-1003.

TORGERSON W. S. (1958). *Theory and Methods of Scaling.* New York : Wiley.

TSIANCO M.C., et GABRIEL K.R. (1984), "Modeling temperature data : An illustration of the use of biplots and bimodels in non-linear modeling." *Journal of Applied Meteorology*, **23**, 787-799.

TUKEY J.W. (1949). "One degree of freedom for non-additivity." *Biometrics*, **5**, 232-242.

UNDERHILL L. G. (1990). "The coefficient of variation biplot." *Journal of Classification*, **7**, 41-56.

VALOIS J.-P. (2000). «Approche graphique en analyse de données». *Journal de la Société Française de Statistique*, 141, 5-40.

van EEUWIJK F. A. (1995). "Multiplicative interaction in generalised linear models." *Biometrics* **51**, 1017-32.

VENABLES W. N. et RIPLEY B.D. (1999). *Modern Applied Statistics*, 3e édition. New York : Springer.

VERBOON P. (1994). *A Robust Approach to Nonlinear Multivariate Analysis.* Leiden, DSWO Press.

WOLD H. (1966). "Nonlinear estimation by iterative least squares procedures." *Research Papers in Statistics* (F. N. David, ed.). New York, Wiley, pp. 411-444.

WOLD S. (1976). "Cross-validatory estimation of the number of components in factor and principal component analysis." *Technometrics*, **20**, 397-405.

YAN W., HUNT L. A., SHENG Q. et SZLAVNICS Z. (2000). "Cultivar evaluation and mega-environment investigation based on the GGE biplot." *Crop Science*, **40**, 597-605.

Appendice : Un algorithme de régression pour la validation croisée des approximations de rang réduit

Pour une matrice $\mathbf{Y}(n \times m)$ et le rang réduit $r \leqslant rang(\mathbf{Y})$ on commence avec l'élément $y_{1,1}$ en créant la division

$$\mathbf{Y} = \begin{bmatrix} y_{1,1} & \mathbf{y}_{1\bullet}^T \\ \mathbf{y}_{\bullet 1} & \mathbf{Y}_{\backslash 1,1} \end{bmatrix}, \tag{A.1}$$

et on approxime la sous-matrice $\mathbf{Y}_{\backslash 1,1}$ par son ajustement de rang r,

$\sum_{k=1}^{r} \widetilde{\mathbf{u}}_{(k)}\widetilde{\mathrm{d}}_k\widetilde{\mathbf{v}}'_{(k)} = \widetilde{\mathbf{U}}\widetilde{\mathbf{D}}\widetilde{\mathbf{V}}^T$ où $\mathbf{Y}_{\backslash 1,1} = \sum_{k=1}^{rang(\mathbf{Y}_{\backslash 1,1})} \widetilde{\mathbf{u}}_{(k)}\widetilde{\mathrm{d}}_k\widetilde{\mathbf{v}}'_{(k)}$ est une décomposition singulière et $\widetilde{\mathbf{U}} = \left[\widetilde{\mathbf{u}}_{(1)} \ldots \widetilde{\mathbf{u}}_{(r)}\right]$, $\widetilde{\mathbf{V}} = \left[\widetilde{\mathbf{v}}_{(1)} \ldots \widetilde{\mathbf{v}}_{(r)}\right]$, $\widetilde{\mathbf{D}} = \mathrm{diag}(\widetilde{\mathrm{d}}_1 \ldots \widetilde{\mathrm{d}}_r)$, $\widetilde{\mathrm{d}}_1 \geqslant \widetilde{\mathrm{d}}_2 \geqslant \ldots \geqslant \widetilde{\mathrm{d}}_{rang(\mathbf{Y}\backslash 1,1)}$. Puis on utilise la régression $\widetilde{\mathbf{U}}\widetilde{\mathbf{D}}^{-1}\widetilde{\mathbf{V}}'\mathbf{y}_{\bullet 1}$ (ou $\widetilde{\mathbf{V}}\widetilde{\mathbf{D}}^{-1}\widetilde{\mathbf{U}}'\mathbf{y}_{\bullet 1}$) de la première ligne (ou première colonne) omettant la première colonne (ou ligne) pour obtenir la « prédiction » de $y_{1,1}$ par

$$\breve{y}_{1,1}^{\{r,0\}} = \mathbf{y}'_{1\bullet}\widetilde{\mathbf{V}}\widetilde{\mathbf{D}}^{-1}\widetilde{\mathbf{U}}'\mathbf{y}_{\bullet 1} \tag{A.2}$$

et le résidu $y_{1,1} - \breve{y}_{1,1}^{\{r,0\}}$.

De la même façon on obtient les « prédictions » $\breve{y}_{1,1}^{\{r,0\}}$ et les résidus $y_{1,1} - \breve{y}_{1,1}^{\{r,0\}}$ pour tous les autres éléments $y_{i,j}$, $i = 1, \ldots, n$; $j = 1, \ldots, m$; $(i,j) \neq (1,1)$. (Chacun utilisera une division différente de $\mathbf{Y}$.)

On a parfois trouvé que ces calculs donnaient de très mauvaises prédictions et que cela arrivait quand certaines valeurs singulières étaient très proches de zéro. On a donc formulé une variante de cette méthode avec un coefficient de ridge c, selon laquelle la prédiction de $y_{1,1}$ devient

$$\breve{y}_{i,j}^{\{r,c\}} = \mathbf{y}'_{1\bullet}\widetilde{\mathbf{V}}(\widetilde{\mathbf{D}} + c\mathrm{d}_1\mathbf{I}_r)^{-1}\widetilde{\mathbf{U}}'\mathbf{y}_{\bullet 1}(1+c), \tag{A.3}$$

et les autres prédictions changent de manière analogue.

On peut mesurer la qualité de l'ensemble de ces prédictions par $corr^2\left(y_{i,j}, \breve{y}_{i,j}^{\{r,c\}} \middle| \forall i,j\right)$ ou en utilisant $\sum_{i=1}^{n}\sum_{j=1}^{m}(y_{i,j} - \breve{y}_{i,j}^{\{r,c\}})^2$ qui est une version du *PRESS* (Allen, 1974).