REVUE DE STATISTIQUE APPLIQUÉE

LE CERCLE FACTORIEL

Exploitation graphique des plans factoriels

Revue de statistique appliquée, tome 45, n° 4 (1997), p. 39-64

<http://www.numdam.org/item?id=RSA_1997__45_4_39_0>

© Société française de statistique, 1997, tous droits réservés.

L'accès aux archives de la revue « Revue de statistique appliquée » (http://www.sfds.asso.fr/publicat/rsa.htm) 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/

Rev. Statistique Appliquée, 1997, **XLV** (4), 39-64

EXPLOITATION GRAPHIQUE DES PLANS FACTORIELS

Le Cercle Factoriel[1]

RÉSUMÉ

Les moyens informatiques graphiques actuels doivent faciliter la tâche d'interprétation des résultats des analyses factorielles des données. Le «Cercle Factoriel», sous groupe du groupe «Logiciel et Statistique» de l'ASU, a travaillé à la définition d'un ensemble de fonctions graphiques qui pourraient figurer dans les futurs logiciels d'exploration des tableaux de données. On distingue les fonctions générales d'habillage des éléments d'un plan factoriel et les fonctions à caractère statistique destinées à faciliter l'interprétation du graphique. Les différents types d'analyses sont détaillés ainsi que la procédure du Biplot.

Mots-clés : *Analyse en composantes principales, analyse des correspondances, analyse factorielle, Biplot, classification, exploration graphique, feuille de style, plan factoriel.*

ABSTRACT

Interpretation of exploratory data analysis results should be made easier using the interactivity of computed graphics. The «Cercle factoriel», a subgroup of the Statistical Computer Group of ASU, defined a set of fonctionalities to be implemented in any data analysis software. General graphical fonctionalities for any element in principal axes graphics are described. Statistical graphical fonctionalities are then specified for interpretation of each kind of analysis. We finaly present the Biplot procedure.

Keywords : *Biplot, clustering analysis, correspondence analysis, exploratory analysis, factorial graphics, principal components analysis, style sheets.*

1. Introduction

Les moyens informatiques actuels permettent d'améliorer considérablement l'exploitation des résultats d'analyse des données. Ceci est rendu possible grâce notamment à un affichage graphique de qualité et à l'interactivité. L'interactivité permet à l'utilisateur de modifier en temps réel les éléments du graphique par l'intermédiaire du clavier ou de la souris.

Le «Cercle Factoriel» – sous-groupe du Groupe ASU-Logiciels et Statistique – s'est fixé comme objectif de réfléchir à l'exploitation graphique des plans factoriels

(1) Membres rédacteurs : D. Ambroise, J-M. Bernard, J-L. Blanchard, F. Goupil-Testu, D. Grangé, A. Morineau, F. Sermier, G. Thauront, N. Valette. Le groupe est animé par A. Morineau (CISIA, Saint-Mandé).

en s'appuyant sur l'expérience de ses participants, en s'inspirant des éléments bibliographiques cités en référence et en faisant l'analyse de certains logiciels existant.

Le «Cercle Factoriel» a restreint son étude et ses propositions au cas des plans factoriels, sachant que les représentations à trois dimensions présentent des problèmes spécifiques. D'autre part ce travail est limité aux plans issus des méthodes classiques suivantes :

- l'analyse en composantes principales, normée ou non (ACP);
- l'analyse des correspondances simples (AFC);
- l'analyse des correspondances multiples (ACM).

On s'appuie ici sur les propriétés classiques de ces analyses dont on trouvera la présentation dans la plupart des manuels sur l'analyse des données. Pour les rappels sur ces méthodes et pour compléter les références bibliographiques, on pourra consulter par exemple [Saporta, 1990; Rouanet *et al.*, 1993; Tenenhaus, 1994; Lebart *et al.*, 1995].

Un exposé plus détaillé de ce travail a fait l'objet d'un rapport dont l'essentiel a été publié dans *La Revue de Modulad*, volume 17, 1996, pp. 31-95 (Edition INRIA, Rocquencourt).

2. Généralités

2.1. Le plan factoriel : un plan très particulier

Un plan factoriel se différencie d'un graphique (x, y) usuel de plusieurs façons :

- Les axes sont «hiérarchisés» : un des axes est plus important que l'autre au sens où il indique une direction de plus grande dispersion (ou inertie ou allongement) que l'autre.
- Les points positionnés sur le graphique peuvent être de différents types (par exemple des points-lignes et des points-colonnes). Ils peuvent jouer des rôles différents : actifs ou illustratifs.
- Les unités sur les axes dépendent dans certains cas de l'intention du graphique.
- La localisation précise des points dans le plan factoriel ne permet pas d'apprécier exactement la distance réelle entre les points ni la distance réelle au centre du graphique.
- Les zones du graphique ne sont pas équivalentes (par exemple le centre du graphique est souvent moins intéressant que la périphérie et les distances ne s'y lisent pas de la même façon).
- Certains points définissent plutôt des directions que des localisations. Dans ce cas on peut répéter pour les angles entre directions ce que l'on vient de dire pour les distances entre points.
- Les distances et les angles se lisent avec des règles de lecture qui diffèrent selon la nature de l'analyse et la nature des points.
- Le sens des axes est arbitraire (toutes les symétries sont possibles).

2.2. Rappel de vocabulaire

2.2.1. Les points-individus

Ils correspondent en général aux lignes du tableau analysé par ACP ou ACM (et dans certaines applications de l'AFC). Il s'agit des «individus statistiques» c'est-à-dire des objets en général extraits d'une population et sur chacun desquels les observations ont été faites.

2.2.2. Les points-variables continues

Un point-variable continue est défini par ses n coordonnées dans l'espace à n dimensions des individus.

En ACP, on définit une distance entre deux variables continues actives. Suivant le cas cette distance s'exprime à partir des covariances (analyse non normée) ou à partir des corrélations (analyse normée). Le nuage des points-variables actives dans un plan factoriel permet de visualiser ce type de distances entre variables. Les points-variables continues illustratifs sont positionnés dans le plan en fonction de leurs corrélations avec les axes factoriels.

Entre deux variables continues illustratives, ou entre une active et une illustrative, il n'y a jamais de calcul de distance.

2.2.3. Les points-anciens axes unitaires

Une analyse en composantes principales peut être considérée comme un changement de repère orthonormé. Le nuage des points-individus est initialement construit dans le repère originel des variables, colonnes du tableau des observations. Un axe unitaire de ce repère donne, dans ce nuage d'individus, la direction dans laquelle la variable correspondante va croissant. Le transfert de ces anciens axes unitaires dans le nouveau repère des axes factoriels définit des directions associées aux variables et dont la position au sein des individus présente le même caractère : la direction dans laquelle la variable va croissant.

2.2.4. Les points-modalités

Ce sont les points représentant les modalités des variables nominales. Ils représentent donc des groupes d'individus. On verra que, dans tous les cas, les coordonnées de ces points sont les moyennes des coordonnées des individus du groupe (éventuellement à un coefficient près dépendant de la dispersion sur l'axe).

2.2.5. Les points-fréquences

Ce sont les colonnes (et généralement aussi les lignes) d'un tableau de fréquences soumis à une AFC. Mais ce sont parfois des colonnes de fréquences dans

un tableau «individus × variables». Dans ce cas ce sont des variables illustratives pour une ACP ou une ACM.

2.2.6. Rôle actif ou illustratif

Un point est actif s'il participe au calcul de l'inertie du nuage de points. Il participe alors à la détermination des axes du plan factoriel (les directions de plus grandes inerties). Sinon il est dit illustratif ou supplémentaire. Le choix du rôle joué par un point est essentiel dans la construction de l'analyse et la connaissance de ce rôle est essentielle au moment de l'interprétation des résultats.

3. Fonctions graphiques

3.1. Introduction

On peut distinguer deux usages des graphiques factoriels :

- les graphiques à des fins exploratoires des données
- les graphiques de production pour publication.

En fait, les fonctions nécessaires à ces deux types de graphiques sont sensiblement les mêmes. L'interactivité nécessaire dans la phase exploratoire l'est aussi pour la mise au point d'un graphique à publier. On peut donc raisonnablement envisager un seul produit dont certaines fonctions seront plus importantes dans la phase exploratoire et d'autres dans la phase de production.

3.1.1. Les fonctions de base

On liste dans ce paragraphe un ensemble de fonctions qui doivent constituer la base de tout logiciel d'exploration des plans factoriels.

Le programme doit permettre de visualiser en permanence le graphique en cours pour apprécier de façon interactive les modifications effectuées. On doit également pouvoir sauvegarder, à tout moment, le graphique et les éléments qui ont permis sa constitution de façon à autoriser la reprise ultérieure du travail pour le terminer ou pour le modifier.

L'utilisateur doit pouvoir définir des modèles de feuilles de style qui seront utilisées ou réutilisées pour tous les plans que l'on désire dessiner.

Dans la phase exploratoire l'interactivité avec les données de base est nécessaire. Il doit être possible en particulier de faire apparaître toutes les informations concernant un point en «cliquant» sur le point (on s'autorisera l'utilisation du verbe *cliquer* bien qu'il soit incorrect en français).

Dans la phase de production, le graphique doit être construit en tenant compte du type de produit que l'on veut obtenir :

- un graphique pour publication papier en noir et blanc,
- un graphique pour publication papier en couleur,

- une sortie sur transparent,
- une sortie sur diapositive,
- une sortie sur écran d'ordinateur.

Il est souhaitable que l'utilisateur puisse imprimer un graphique «par morceaux» (impression multi-pages).

Un graphique achevé pour un type de sortie (transparents couleurs par exemple) doit pouvoir être transformé automatiquement pour un autre type de sortie (publication papier noir et blanc par exemple).

En fonction du type de support (sur papier noir et blanc, sur papier couleur, transparents, diapositives), le logiciel proposera un graphique par défaut en faisant des choix adaptés au type de sortie. Les problèmes de recouvrement des points auront été résolus automatiquement.

3.1.2. La notion de groupes de points

Les différents points d'un plan peuvent être regroupés pour faire des traitements communs au niveau de leur représentation, de leur sélection ou désélection. Les groupements à envisager sont :

- les variables continues actives ou illustratives
- les modalités actives ou illustratives
- les individus actifs ou illustratifs
- les individus en fonction de l'appartenance à une classe (suite à une classification)
- les individus en fonction des modalités d'une variable.

Il y a autant de groupes d'individus qu'il y a de classes ou de modalités. Il peut être intéressant «d'éclater» le plan factoriel en autant de fenêtres graphiques que de modalités ou classes. Dans le cas d'une classification des individus, ou d'une représentation suivant les modalités d'une variable, un autre groupe apparaît : celui des centres de gravité des classes d'individus.

Enfin on peut avoir besoin de définir des groupes de points, sélectionnés suivant des critères statistiques (cosinus carré, contribution, etc.).

3.1.3. Représentation des variables

Les points-variables d'un graphique factoriel doivent être affectés d'un symbole et d'un libellé. Le symbole doit être placé aux coordonnées exactes du point. Le libellé doit être placé le plus près possible du symbole correspondant. Le logiciel doit calculer la position de ce libellé en tenant compte de son environnement et en évitant les recouvrements de libellés. Le libellé doit être mobile autour du symbole. La possibilité de mettre une flèche qui pointe du libellé vers le symbole permettra de clarifier le graphique si ce libellé est trop éloigné.

Les attributs de style des libellés des variables doivent pouvoir être traités point par point ou pour tout un groupe. Le texte du libellé et ses attributs doivent être modifiables à tout moment.

3.1.4. Représentation des individus

On peut vouloir représenter le nuage des individus de différentes façons :

- par les identifiants ou libellés courts : on distinguera les individus actifs des illustratifs par des polices de caractères et/ou des couleurs différentes,
- par des symboles, différents selon le rôle actif ou illustratif,
- selon les valeurs d'une variable continue,
- selon les modalités d'une variable nominale ou les classes d'une partition : les individus seront représentés par des symboles différents et/ou des couleurs différentes selon les groupes. On distinguera aussi les individus actifs des illustratifs.

On aura la possibilité de faire apparaître l'identifiant de l'individu en cliquant sur le point. Cette fonction est intéressante pour mettre un individu particulier en évidence.

Dans le cas d'un trop grand nombre d'individus on représente les seuls centres de gravité des classes et des modalités. Ces centres de gravité seront traités comme des points illustratifs et dotés de symboles et de libellés. Il peut être intéressant de mettre les symboles des centres de gravité dans une taille proportionnelle à l'importance de la classe (voir la Figure 1). Il pourra aussi être intéressant de relier (par des segments) chaque centre de gravité à l'ensemble des points de sa classe, ce qui permet de visualiser la dispersion de la classe (voir la Figure 7).

3.2. Fonctions d'habillage des éléments

3.2.1. Habillage des libellés

On aura le choix entre des libellés courts, longs, ou la concaténation des deux, aussi bien pour les variables continues que pour les modalités. Si plusieurs points sont superposés, on doit pouvoir afficher lisiblement les libellés près du symbole associé.

Attributs des libellés

Ils doivent être contrôlés par :

- le choix des polices de caractères, auxquelles il sera possible d'associer des attributs (gras, italique, souligné).
- le choix des couleurs. Huit couleurs franches sont suffisantes. On notera les difficultés liées aux couleurs; par exemple, le jaune est visible sur un écran ou une diapositive sur fond noir mais n'est pas visible sur papier blanc ou sur transparent.

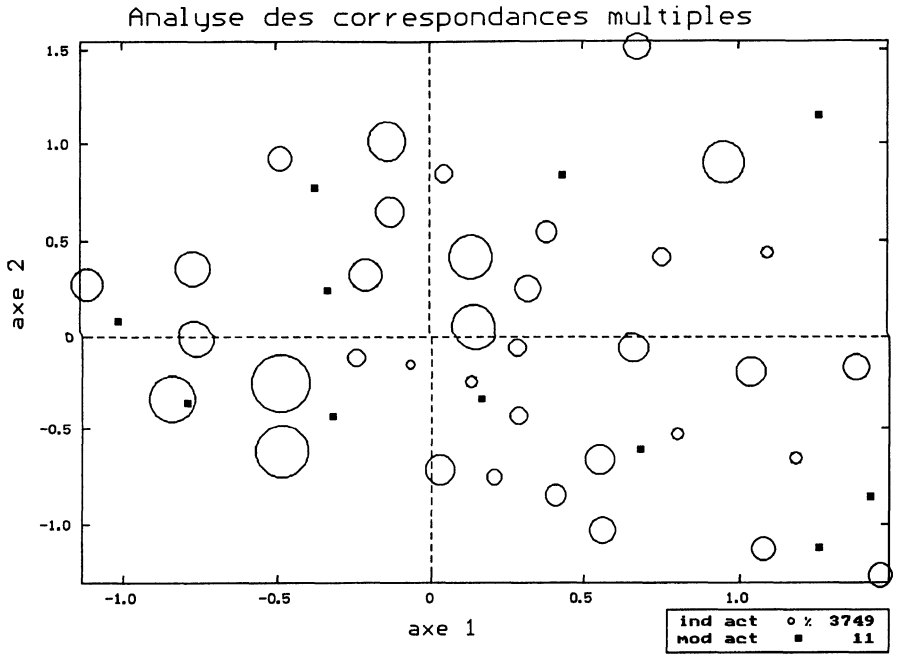

FIGURE 1

Traitement des points superposés.
On repère les points superposés en leur affectant un symbole circulaire dont la surface est proportionnelle au nombre de points superposés (SPAD).

- le choix de la taille des polices de caractères. Proposé en % de l'écran, il permettra de conserver les proportions du graphique sur imprimante.

L'effacement des libellés

Les libellés seront effaçables (et réaffichables) individuellement, ou globalement à l'intérieur ou à l'extérieur d'une zone délimitée par l'utilisateur, ou pour tout un groupe.

3.2.2. Les symboles

On doit avoir le choix de la forme, la couleur et la taille des symboles. L'utilisateur doit pouvoir aussi affecter un symbole à toute sélection de points.

Attributs des symboles

- Le choix des symboles devra être fait en proposant une liste de symboles dont on peut contrôler la taille et la couleur. Aux symboles usuels : carré, cercle,

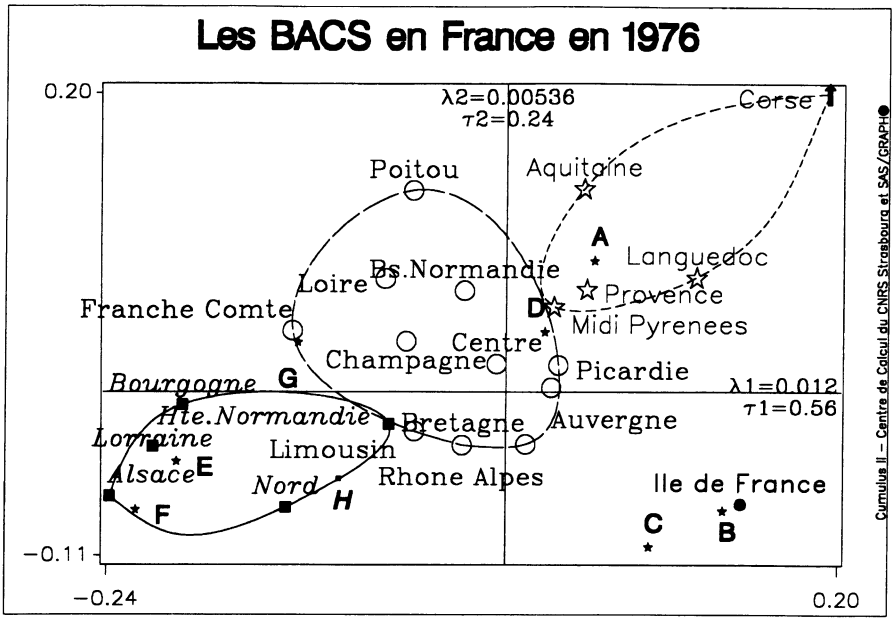

FIGURE 2
Les Bacs en France (Cumulus)

triangle, étoile, losange peuvent être ajoutés des symboles associés à certaines polices spéciales.

- Le choix des couleurs sera fait en proposant une palette de couleurs.
- La taille des symboles pourra être proposé en % de l'écran.

L'effacement des symboles

Les symboles seront effaçables individuellement ou globalement à l'intérieur ou à l'extérieur d'une zone délimitée par l'utilisateur, ou pour toute sélection. Les points sont alors représentés par leurs seuls libellés.

3.2.3. Les flèches

Des flèches seront utilisées :

- pour marquer les points qui sont ramenés sur le bord du cadre (avec une longueur qui peut être proportionnelle à la distance du point au cadre);
- pour pointer la position exacte d'un point quand le libellé est trop éloigné (elle pourrait apparaître automatiquement dès que la distance atteint une certaine valeur);

- pour représenter des anciens axes unitaires en ACP par exemple (voir la Figure 3).

  On pourra contrôler les valeurs et attributs des flèches.

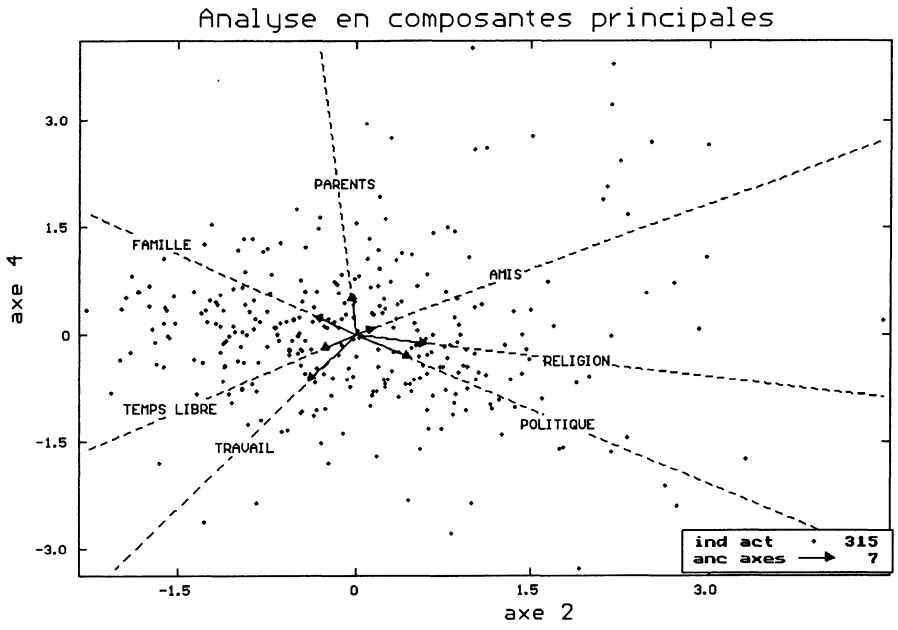

FIGURE 3

Représentation simultanée en ACP (SPAD)

Les flèches représentent les vecteurs unitaires sur les directions des anciens axes.

3.2.4. Le tracé de lignes

Deux types de lignes peuvent être nécessaires :

- les segments de droites pour indiquer des trajectoires entre éléments de même nature; ou pour joindre des modalités d'une même variable (voir la Figure 4); pour joindre le centre de gravité d'une classe aux individus appartenant à cette classe (voir la Figure 7).
- Les courbes pour entourer un groupe de points (voir la Figure 2).

Attributs des lignes

- Le type de ligne est choisi dans une liste de lignes disponibles : ligne pleine, pointillée, tirets longs
- L'épaisseur de la ligne pourra être proposée en % de l'écran.
- La couleur de la ligne sera accessible via une palette de couleurs.

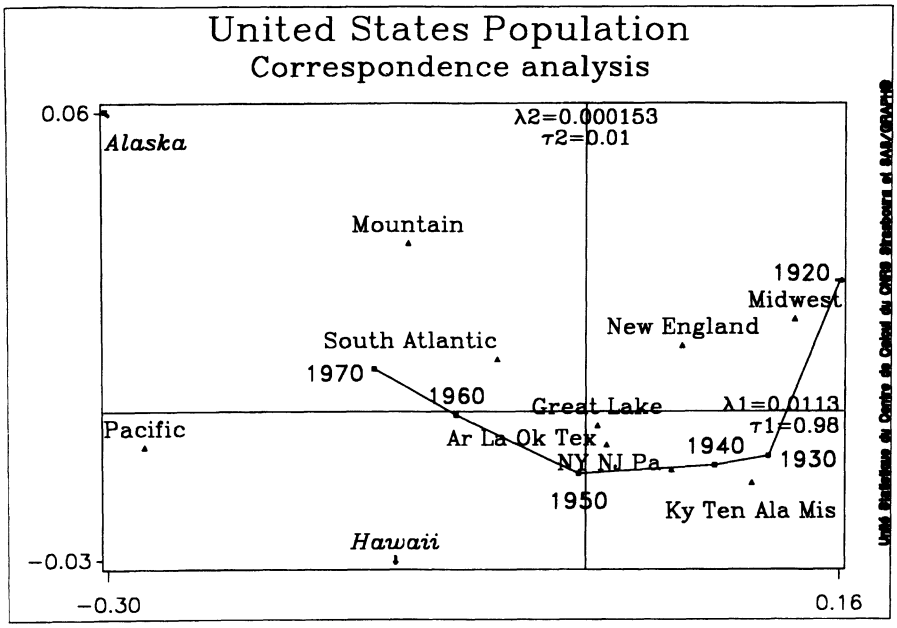

FIGURE 4
Analyse des correspondances (Cumulus)

3.3. Manipulation des éléments

3.3.1. La sélection individuelle

Cette fonction est nécessaire pour permettre la sélection (ou la désélection) des divers points du plan. En particulier pour :

- identifier individuellement chaque point du plan par un simple clic de la souris.
- afficher des informations concernant un élément du plan, par exemple ses coordonnées sur les axes, sa contribution aux axes, la qualité de sa représentation sur les axes.

3.3.2. La sélection d'un groupe

La sélection (ou désélection) d'un groupe de points se fera de plusieurs façons soit selon la nature du groupe, soit à partir d'un critère. On peut vouloir par exemple sélectionner :

- un groupe de points situés à l'intérieur (ou à l'extérieur) d'une zone délimitée par l'utilisateur;
- un groupe d'éléments actifs ou un groupe d'éléments illustratifs (variables ou individus);
- une classe d'individus associés à une modalité d'une variable nominale;

- une classe de points obtenue par une méthode de classification;
- le groupe des centres de gravités des classes;
- un groupe de points selon un critère de plus forte contribution au plan;
- un groupe de points selon un critère de bonne représentation sur ce plan.

3.3.3. Le zoom

La fonction zoom sera disponible avec indication visible des bornes. La zone à agrandir doit être délimitée par l'utilisateur et réaffichée en conservant toutes les informations qu'elle contenait antérieurement : par exemple, identifiants d'éléments et segments.

Le zoom doit pouvoir être utilisé «en cascade». Le niveau de zoom sera alors indiqué à l'écran.

Les limites du zoom seront mémorisées à chaque étape pour permettre le retour arrière par les mêmes étapes. On doit pouvoir revenir à l'état initial du graphique à partir de n'importe quel niveau de zoom. Enfin, toute manipulation qui cache des points donnera lieu à l'ouverture d'une fenêtre situant la position de la zone agrandie à l'intérieur du graphique.

3.3.4. Un zoom particulier

Un zoom particulier permet de se concentrer sur la partie la plus dense du graphique tout en conservant l'ensemble des points. Ce zoom particulier est obtenu de la façon suivante : les points extérieurs à la zone dense sont ramenés à la périphérie du graphique dans leur direction par rapport à l'origine (procédure appelée «zoom ourlé» dans SPAD). Le zoom ordinaire doit être disponible à l'intérieur de ce zoom particulier.

3.3.5. Étirement

Le rectangle contenant le graphique sera étirable dans le sens de la longueur ou de la largeur dans la limite de la place disponible à l'écran.

3.3.6. Graduations

Le choix du type de graduation sur le graphique sera fait selon que l'on souhaite avoir des échelles identiques ou utiliser au mieux la surface de l'écran. Pour ne pas charger le graphique, il est préférable que les graduations figurent sur le cadre.

En analyse en composantes principales, on pourra faire apparaître sur les directions des anciennes variables des graduations correspondant aux unités des variables.

3.4. Fonctions d'habillage du plan

3.4.1. Les axes

On doit pouvoir faire apparaître, à proximité des axes, leur numéro, les valeurs propres associées, les taux d'inertie de ces axes (voir la Figure 2).

3.4.2. Le cadre

Le cadre entourant le graphique sera affichable ou non et on aura le contrôle de l'épaisseur et de la couleur de ce cadre. Les graduations pourront figurer ou non sur ce cadre.

3.4.3. Le quadrillage

Faire figurer un quadrillage sur le graphique peut être utile pour faciliter les repérages.

3.4.4. L'ajout de texte ou de titre

Il est intéressant de pouvoir :

- ajouter un titre ou un texte à un endroit quelconque du graphique. Ce titre ou texte doit pouvoir être affiché horizontalement ou verticalement avec contrôle de la police de caractères, la couleur et la taille.
- modifier, déplacer ou effacer un texte.

3.4.5. La légende

Dès que l'on donne des significations sémantiques à des éléments graphiques, il faut les expliciter dans des cartouches de légendes déplaçables et éditables (voir la Figure 3)

Exemple de légende explicative :

«*Les individus sont représentés par des symboles dépendant de la modalité de la variable x; la couleur du symbole dépend de la modalité de la variable y; les variables illustratives sont en italique*».

La légende pourra contenir des informations statistiques (les effectifs par exemple).

3.4.6. La définition et coloriage d'une zone

On pourra entourer une zone de points à l'aide de la souris avec possibilité d'effacement du tracé et contrôle de la couleur, de l'épaisseur et du type de trait. On pourra colorier ou choisir une trame pour une zone fermée.

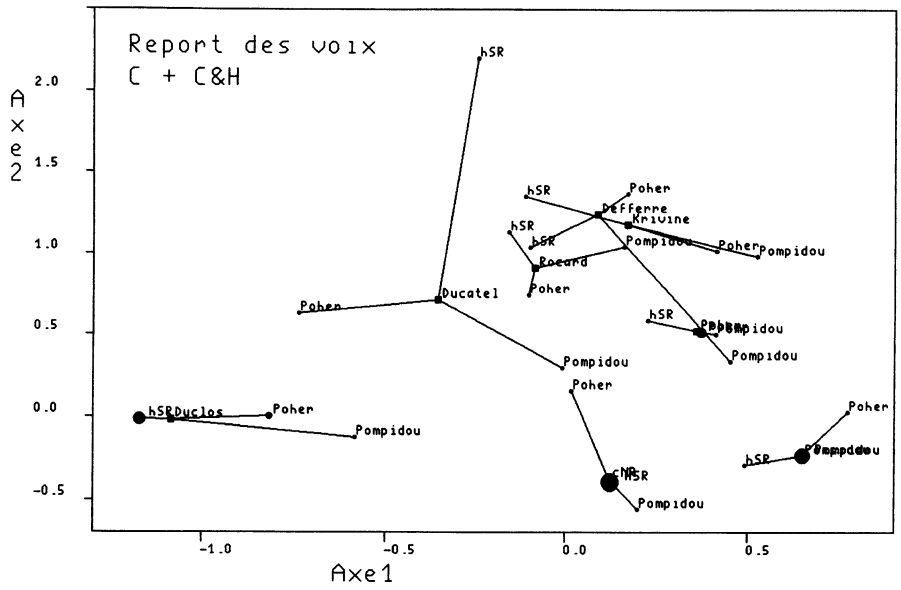

FIGURE 5
Segments dans une analyse des correspondances (EyeLID)

3.5. Styles, feuilles de styles et modèles

3.5.1. Styles et feuilles de styles

À chaque élément du graphique est associé un ensemble d'attributs qui peuvent être purement graphiques ou qui présentent un caractère statistique.

La notion de style permet de regrouper l'ensemble de ces attributs et de leur donner un nom. Il est possible de sauvegarder le style soit dans le fichier contenant le graphique, soit dans une feuille de style autonome, soit dans un fichier d'options. On peut ainsi récupérer les définitions d'attributs et les appliquer aux éléments d'un autre graphique de la même analyse ou plus généralement à une autre analyse.

3.5.2. Les graphiques modèles

Les choix adoptés pour la réalisation d'un graphique doivent pouvoir être stockés de manière indépendante des données. On peut envisager de stocker ces informations dans des fichiers de modèles indépendants les uns des autres ou de les regrouper dans une bibliothèque.

L'utilisateur peut créer de nouveaux modèles, modifier des modèles existants, les supprimer et bien sûr, les appliquer à ses analyses, tout en gardant la faculté de modifier les différents attributs.

On peut envisager de fournir un jeu de modèles adaptés :

- aux différentes analyses factorielles (ACP, AFC, ACM)
- dans des contextes particuliers
  - petit nombre de points ou très grand nombre de points
  - modalités qualitatives ordonnées,...
- pour différents supports
  - papier, transparent, écran, diapositive,
  - noir et blanc ou couleur.

3.5.3. Exemples de réalisation

Nous évoquons quelques réalisations dans différents logiciels du commerce : SAS, JMP et Excel.

Dans le module SAS/GRAPH de SAS Institute, il est possible de gérer une liste de définition d'axes, de motifs ou de symboles (instructions AXIS, PATTERN, SYMBOL,...). Ils peuvent être définis par programme ou de manière interactive dans des fenêtres spécialisées. Ces définitions sont ensuite utilisées dans les instructions qui génèrent les graphiques. Enfin, de nombreuses options graphiques permettent de choisir les paramètres graphiques adaptés aux différents périphériques de sortie. Le travail de l'utilisateur de SAS/GRAPH consiste à déterminer un choix de paramètres adapté à ses besoins. Le travail consiste à gérer une description (programme SAS) d'un ensemble de styles et de modèles de graphiques.

Le logiciel JMP sur Apple Macintosh (et d'une manière similaire, le module SAS/Insight) propose une approche très différente. Les attributs graphiques et statistiques des points sont des éléments associés à chaque ligne de la table de données et il est possible de les stocker, avec la table, comme autant de variables nouvelles. Le point fort, sous l'angle qui nous intéresse ici, est la facilité de créer l'habillage d'un graphique et la grande facilité de stocker cet habillage et de le réutiliser. Cette facilité ne s'applique qu'à l'intérieur d'un même tableau de données. Il n'est pas possible de définir des styles au sens où nous les avons définis, pas plus que de créer des documents types.

Dans le tableur Excel de Microsoft, il est possible de définir des styles qui sont attachés à chaque document. Il est possible de récupérer en bloc l'ensemble des styles définis dans un document et de les transporter dans un autre document. Il est également possible de créer des documents modèles. Enfin, on peut définir des graphiques types, repérés par un nom et par une description sommaire. Ils sont gérés dans un fichier d'options (ou de préférences) et il est facile d'appliquer le graphique type à toute autre série de données.

3.5.4. Exemple de dialogue avec Excel

La première image (Figure 6.1) montre le dialogue de définition du style d'un point réalisé à partir de Excel. La définition du motif (trait et marque) débouche sur un dialogue visible sur la Figure 6.2. La Figure 6.3 montre un exemple de plan factoriel avec deux styles de points : individus actifs et individus illustratifs.

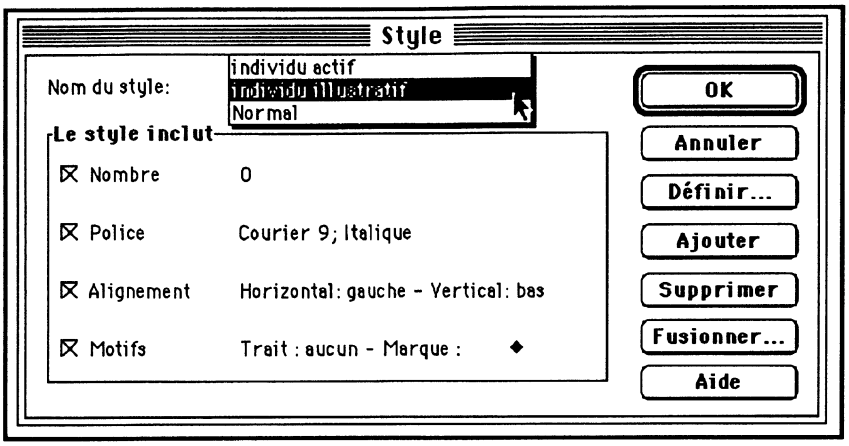

FIGURE 6.1
Exemple de dialogue de définition de style (programmation Excel)

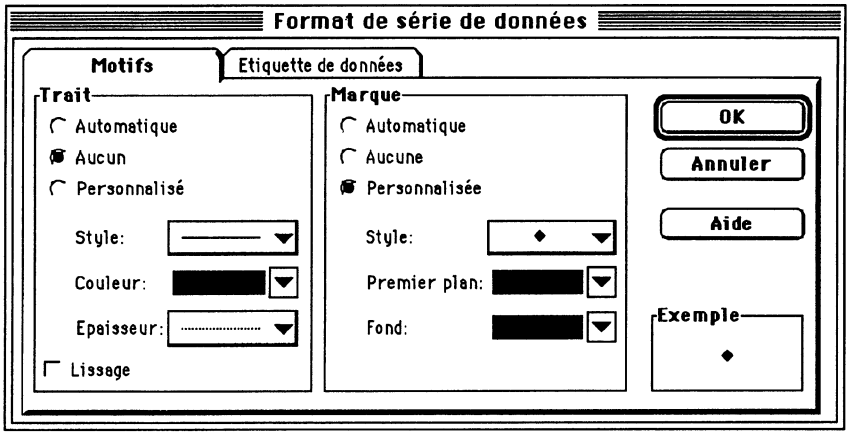

FIGURE 6.2
Exemple de dialogue de définition de style (suite)

4. Fonctions statistiques

4.1. Fonctions statistiques selon le type de points

4.1.1. Généralités

Il y a peu de fonctions liées à la nature de l'analyse (ACP, ACM ou AFC) sinon les règles de représentations simultanées. Par contre des fonctions importantes sont liées à la nature des points : individus, variables continues, modalités ou classes, fréquences. On différenciera points actifs et illustratifs car les règles d'interprétation

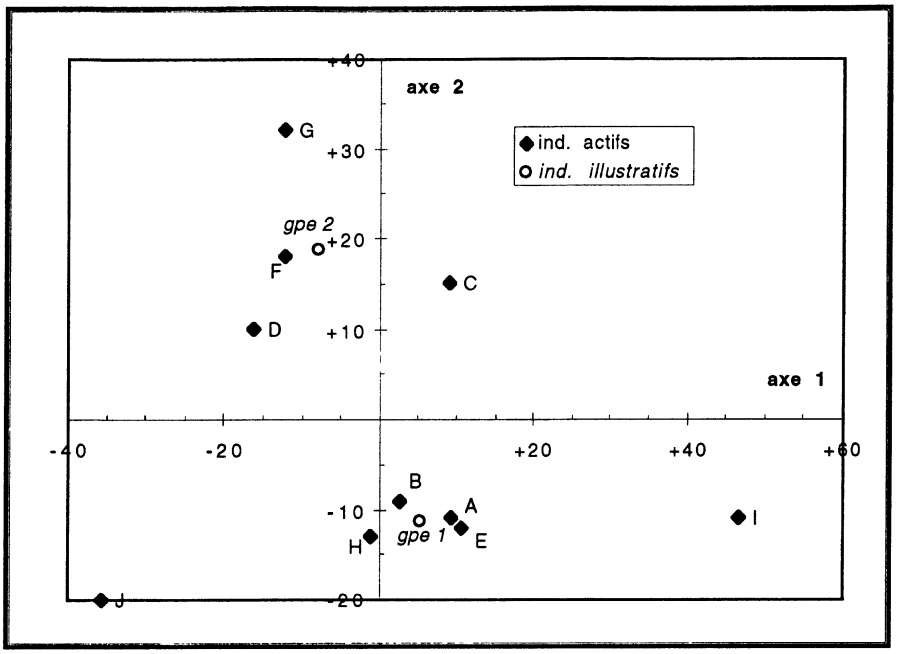

FIGURE 6.3
Exemple de graphique factoriel (programmation Excel)

des proximités ne sont pas les mêmes (même si les calculs des positionnements sont identiques).

Chaque fois que l'on calcule une distance ou une inertie (ou une quantité dérivée comme une contribution, un $\cos^2$, un plus proche voisin, etc.) pour habiller un graphique, il faut spécifier le sous espace servant de support au calcul : soit un seul axe, soit le premier plan, soit l'espace jusqu'à l'axe de plus haut rang du plan représenté soit l'espace complet. Il y aura souvent un sous-espace par défaut (par exemple : on calcule les k plus proches voisins dans l'espace complet pour les identifier sur le premier plan factoriel).

S'il y a sélection de points selon un critère continu (ex : la contribution), on actionnera des «jauges» ou des curseurs. En général les points s'estomperont mais on pourra garder un «fantôme» de la position des points non sélectionnés (un pixel résiduel). Il en sera de même pour une sélection logique (une classe, un filtre logique plus général). En effet il y a souvent intérêt à garder une image de l'ensemble du nuage des points au sein duquel on veut mettre en valeur un sous-ensemble de points.

4.1.2. Pointeur d'information

En pointant sur un point on obtient toutes les informations sur l'élément. Pour un individu, on obtient ses valeurs dans le tableau de données, son poids, ses coordonnées,

contributions et $\cos^2$ sur les axes factoriels, sa distance au centre, sa participation à l'inertie globale, ses classes d'appartenance dans les partitions disponibles.

Pour une variable continue, on obtient sa moyenne, son écart-type, son minimum, son maximum, le nombre de données manquantes, son histogramme, ainsi que ses coordonnées, contributions et $\cos^2$ sur les axes factoriels. Noter qu'un axe est une variable particulière.

Pour une variable «effectifs», on obtient sa moyenne, son écart-type, son minimum, son maximum, son histogramme, ainsi que ses coordonnées, contributions et $\cos^2$ sur les axes factoriels.

Pour une variable nominale (ou une partition, qui est une variable nominale particulière), on obtient l'effectif (et le poids) de chaque modalité, les coordonnées, contributions et $\cos^2$ sur les axes factoriels, ainsi que les valeurs-tests sur les axes.

4.1.3. Cadre et échelles

Il y a deux types de représentation : la forme classique des plans (x, y) avec les axes en bordure et la forme «plan factoriel» avec les axes se croisant au centre de gravité. Cette seconde présentation sera fournie par défaut. On pourra passer de l'une à l'autre.

Noter le problème particulier du nuage des variables continues en ACP (en particulier non normée) : il est souvent indispensable d'imposer l'origine dans le graphique. De plus il est intéressant de «centrer le graphique» dans un cercle de corrélation dans le cas d'une analyse normée même si le nuage des points est situé d'un seul côté de l'origine.

Le choix des échelles et des unités sur les axes factoriels est un choix important. Un cercle de rayon unité pourra être déformé et prendre la forme d'une ellipse pour rendre plus lisible un plan factoriel.

Le sens des axes d'une analyse étant arbitraire, on pourra faire une symétrie horizontale ou verticale pour des raisons esthétiques ou pour comparer plusieurs graphiques.

4.1.4. Mise en valeur de groupes de points

Une opération statistique importante sur le plan factoriel est la sélection de points opérée pour guider l'interprétation du graphique : reconnaître la spécificité des zones du plan, les caractéristiques des directions factorielles, estomper ou effacer les points qui encombrent et perturbent la lecture, etc.

Les principaux critères de sélection sont la contribution à l'inertie (sur un axe, sur le plan, jusque sur le plan), le $\cos^2$ (sur un axe, sur le plan, jusque sur le plan), la distance à l'origine, le poids. Pour les points représentatifs d'un groupe d'individus (modalité ou classe), on peut ajouter la sélection en fonction de la valeur-test au sens défini dans SPAD [Morineau, 1984].

Soit la sélection s'opère en oui/non, soit on procède à un habillage qui met en valeur les points en fonction de la valeur du critère retenu. L'habillage peut porter sur l'écriture du libellé ou sur le symbole qui localise le point. Si les points sont représentés par des petits cercles, on jouera sur le diamètre des cercles. En agissant sur un curseur, on pourra estomper progressivement les points en fonction du critère choisi.

Pour les points illustratifs, on affichera la valeur-test et l'on éliminera à l'aide d'un curseur les points illustratifs dont la valeur-test est inférieure à un seuil de façon à ne garder que les points les plus significatifs. A l'inverse on fera apparaître successivement les points illustratifs les plus significatifs en utilisant la valeur-test.

4.1.5. Connexions entre points

Relier deux points par une ligne introduit souvent une aide à la lecture du graphique. On pourra par exemple relier tous les points d'une classe au centre de la classe (graphique en étoile) ce qui permet de juger de la dispersion de la classe dans le plan (voir la Figure 7).

Dans le même esprit, on peut relier le centre de classe à ses k plus proches voisins par un graphique en étoile. Une autre façon de procéder est de fixer le rayon d'une sphère et de marquer (ou mettre en surbrillance) les points intérieurs à la sphère centrée sur le point.

Pour des points-catégories d'une variable *ordinale* on pourra relier les points dans l'ordre naturel pour dessiner une *trajectoire* qui facilitera l'interprétation. S'il n'y a pas d'ordre entre les catégories, on pourra tracer le *chemin de longueur minimale* (image des proximités réelles) ou le *chemin de longueur maximale* (dispersion des catégories dans le plan).

4.1.6. Exploration d'une classification

À partir d'un centre de classe, on pourra, à l'aide d'un bouton curseur, relier successivement par des segments le centre de gravité aux points les plus proches (au sens de la distance réelle et non de la proximité apparente sur le plan factoriel). À chaque étape la fonction identification des points doit être disponible. Il est intéressant également de repérer les points les plus éloignés d'un centre de classe, c'est-à-dire : relier au centre de gravité d'abord les points les plus éloignés puis les points de plus en plus proches.

4.1.7. Cercles, ellipses, enveloppes

Autour d'un point représentant un groupe d'individus, on peut tracer des «ellipses de confiance» (voir par exemple [Saporta *et al.*, 1986]). On peut aussi tracer l'enveloppe convexe des points projetés dans le plan (voir la Figure 2).

Dans une analyse des correspondances simples, on peut tracer autour de chaque point un cercle dont le rayon mesure d'une certaine façon de combien le profil correspondant diffère du profil moyen, centre du graphique (voir Lebart *et al.*, 1995).

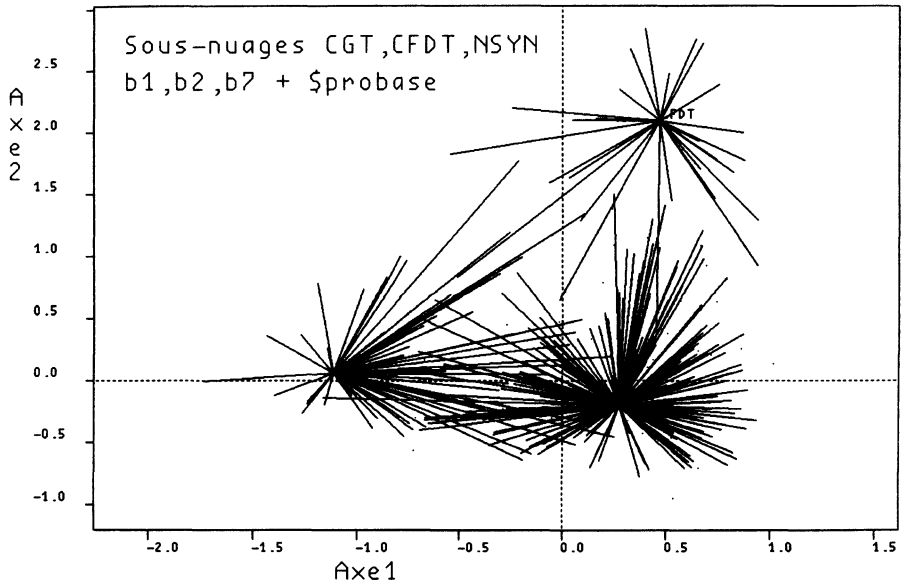

FIGURE 7
Nuages de points moyens et dispersion des classes (EyeLID)

4.1.8. Cas de l'analyse des correspondances simples

La représentation simultanée des lignes et des colonnes de cette analyse est particulière : chaque point d'un nuage étant «quasi barycentre» de l'ensemble des points de l'autre nuage. Cette représentation simultanée sera fournie par défaut. Cependant on devra pouvoir obtenir à la demande le positionnement des lignes en vrais barycentres des colonnes et vice versa.

4.2. Fonctions liées au type «actif/illustratif»

4.2.1. Généralités

La distinction actif/illustratif est essentielle et se traduira par des fonctions graphiques particulières. Il faut rappeler en effet que, entre deux points actifs, il existe toujours une distance interprétable. Par contre entre deux points illustratifs – ou entre un point illustratif et un point actif – il n'y a pas de calcul direct de «distance» et par conséquent les proximités ont une interprétation d'un autre type.

Il est donc essentiel de bien distinguer les points actifs des points illustratifs sur un même graphique ou sur des graphiques proches. La distinction peut se faire par une typographie différente pour les libellés ou par des symboles différents ou par la couleur. On pourra utiliser les attributs italique, gras, souligné, et/ou couleur pour habiller différemment les étiquettes des points.

4.2.2. Superposition des plans

Il est intéressant de disposer séparément du plan «actif» et des différents plans de points illustratifs. Il est intéressant aussi de disposer des représentations simultanées.

On peut imaginer que ces différents graphiques existent sur des supports transparents et sont superposables donc dessinés avec les mêmes échelles. Ces opérations de superposition (avec adaptation automatique des échelles et repositionnement des étiquettes) devraient être réalisables facilement.

4.2.3. Cas de l'analyse des correspondances multiples

Dans une analyse des correspondances multiples, chaque individu représente une certaine combinaison des modalités des variables actives. Lorsqu'il y a beaucoup de points individus et peu de configurations de modalités actives réellement observées, il y aura beaucoup de points-individus confondus. Il est intéressant alors de faire apparaître graphiquement l'abondance des points multiples par exemple en faisant varier le rayon du cercle symbolisant la position de chaque point multiple (voir la Figure 1).

4.2.4. Cas de l'analyse en composantes principales

Considérons la représentation simultanée des individus et des variables en ACP. Les variables sont en fait des individus artificiels représentant les extrémités des anciens axes unitaires porteurs des variables actives. En projection sur un plan factoriel, on obtient des points à l'intérieur d'un cercle unitaire. Mais il est clair qu'il n'y a aucune commune mesure entre les coordonnées des individus réels et les coordonnées de ces individus artificiels que sont les vecteurs unitaires. Pour rendre lisible la figure, on sera amené en général à procéder à une dilatation de l'un des deux nuages.

On prendra garde au fait que dans cette représentation simultanée très particulière des individus et des variables, il n'est pas licite de faire apparaître des variables illustratives (elles ne font pas partie du repère orthonormé des variables actives).

Le nuage proprement dit des variables est une représentation graphique de la matrice des corrélations (ou des covariances si l'analyse n'est pas normée). Dans ce nuage la distance entre deux points s'exprime en fonction de la corrélation (ou des variances et covariance) et se lit en terme d'angles. On prendra donc soin de rendre non superposables tout nuage des individus avec ce nuage des variables. Ici par contre les variables continues illustratives peuvent figurer en même temps que les variables actives.

4.2.5. Cas de l'analyse des correspondances simples

Un point étant un profil (un histogramme de répartition), on doit pouvoir faire apparaître cet histogramme à la demande.

Noter que l'on doit pouvoir faire figurer sur tout plan factoriel d'une AFC des variables continues illustratives et des variables nominales illustratives.

4.3. Représentation simultanée par Biplot

Le Biplot, développé par Gabriel (1971), consiste à faire une approximation de la matrice des données par un produit matriciel de dimension 2 afin de conduire à une représentation plane. La technique de décomposition est la décomposition en valeurs singulières (valeurs propres et vecteurs propres), identique à celle de l'analyse en composantes principales. Le Biplot, comme l'analyse en composantes principales, est généralement utilisé en centrant et réduisant les variables.

La technique du Biplot superpose sur un même graphique les individus et les variables selon trois types de représentation simultanée :

- Soit dans l'espace des variables : *Le cosinus de l'angle entre deux variables approxime la corrélation entre ces variables. La distance entre les individus approxime la distance de Mahalanobis (et non la distance euclidienne de l'ACP).*
- Soit dans l'espace des individus : *La distance entre les individus approxime la distance euclidienne mais la distance entre variable n'est pas interprétable.*
- Soit dans un espace intermédiaire : *Les distances entre individus et entre variables ne sont pas interprétées, mais on obtient un graphique «équilibré».*

Pour l'essentiel, les commentaires graphiques concernant l'ACP s'appliquent donc aux représentations du Biplot. La méthode étant relativement peu utilisée dans le milieu statistique francophone, nous rappelons ses propriétés et ses liens avec l'ACP en annexe.

5. Conclusion

Les fonctions décrites ici peuvent exister dans certains logiciels dévolus aux méthodes d'analyses de données mais il est probable qu'aucun n'en possède l'ensemble (la version pour Windows du logiciel SPAD s'est déjà inspiré de ce «cahier des charges» en cours de développement et intègre les fonctions considérées ici comme essentielles). Cette réflexion sur un «cahier des charges» ne constitue en fait qu'une première étape et ne se prétend pas définitive. Les artisans du «Cercle Factoriel» espèrent plutôt contribuer à un mouvement d'enrichissement mutuel en instaurant une collaboration plus étroite entre les statisticiens (scientifiques et appliqués) d'une part, et les développeurs de logiciels d'autre part.

Références bibliographiques

ALEVISOS P., MORINEAU A. (1992) *Tests et valeurs-tests*. Revue de Statistique Appliquée, Vol.40, n° 4, pp. 27–43.

American Statistical Association (1992, 1993) *Proceedings of the Section on Statistical Graphics.*

BECKER R.A., CLEVELAND W.S., WILKS A.R. (1987). *Dynamic Graphics for Data Analysis*. Statistical Science, Vol 2 n° 4, pp. 355–395.

BERNARD J.-M., ROUANET H. et BALDY R., (1993) – «*EyeLID-2, Manuel de référence et Guide de l'utilisateur*», INDIA, Paris.

CLEVELAND W.S. (1993) *A Model for Studying Display Methods of Statistical Graphics*. Journal Comput. Graphical Stat., Vol 2 n° 4, pp. 323–343.

Computing Science and Statistics (1992) *Graphics and Visualization*. Proceedings 24th Symposium on the Interfaces.

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

GOUPIL-TESTU F. (1995) *Un Outil Graphique Interactif d'Aide à l'Interprétation de Résultats d'Analyse de Données*. Revue MODULAD, n° 15.

GOWER J.C., HAND D.J. (1996) *Biplots*. Chapman & Hall, Londres.

GRANGÉ D., RINGENBACH M. (1991) *Cumulus II ou le Nuage des Analyses Factorielles vu par SAS*. Club SAS, Cannes.

LEBART L., MORINEAU A., PIRON M. (1995), *Statistique exploratoire multidimensionnelle*. Dunod, Paris.

MORINEAU A. (1984), *Note sur la caractérisation d'une classe et les valeurs-tests*. Bull. Techn. Centre Stat. et d'Infor. Appl., vol. 2, pp. 20–27. CISIA, Saint-Mandé.

ROUANET H., LE ROUX B. (1993), *Analyse des données multidimensionnelles*, Dunod, Paris.

SAPORTA G. (1990), *Probabilités, analyse des données et statistique*. Editions Technip, Paris.

SAPORTA G., HATABIAN G. (1986), *Région de confiance en analyse factorielle*. In : Data Analysis and Informatics, Vol. 4, North Holland, Amsterdam, pp. 499–508.

TENENHAUS M. (1994), *Méthodes statistiques en gestion*, Dunod Entreprise, Paris.

WEIHS H. SCHMIDLI (1990) *Online Multivariate Graphical Analysis : Routine Searching for Structure*. Statistical Science, Vol. 5, n° 2, pp. 175–226.

Annexe

La méthode Biplot et ses liens avec l'ACP

La méthode repose sur la reconstitution de la matrice initiale à l'aide du graphique; autrement dit, la projection des individus sur les variables respecte la répartition des données initiales pour cette variable.

Toute matrice Y peut être décomposée en

$$Y = AB'$$

$$(n, p) = (n, k) \times (k, p) \text{ avec } k = \text{rang de } Y$$

Le Biplot consiste à approximer Y par :

$$Y \approx AB'$$
$$(n,p) = (n,2) \times (2,p)$$

Dans cette écriture, A est la matrice des individus et B est la matrice des variables. Sur un graphique Biplot, on représentera les individus par des points et les variables par des vecteurs.

Pour trouver une telle décomposition, on utilise la décomposition en valeurs singulières :

$$Y = U\Lambda V'$$

où V rassemble les vecteurs propres de $Y'Y$. On utilise la matrice diagonale des valeurs propres

$$\Lambda = \begin{bmatrix} \sqrt{\lambda_1} & 0 & 0 & 0 \\ 0 & \sqrt{\lambda_2} & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \sqrt{\lambda_r} \end{bmatrix}$$

où r est le rang de Y. On a :

$$U = YV\Lambda^{-1}$$

En ne gardant que les 2 premières valeurs propres, on approxime Y par :

$$Y \approx \widehat{Y} = U_{[n,2]}\Lambda_{[2,2]}V'_{[2,p]}$$

On peut alors définir différentes approximations de Y en AB'

	A (individus)	B' (variables)
GH' (espace des variables)	U	$\Lambda V'$
Symétrique	$U\Lambda^{1/2}$	$\Lambda^{1/2}V'$
JK' (espace des individus)	$U\Lambda$	V'

GH' : la décomposition est effectuée dans l'espace des variables

JK' : la décomposition est effectuée dans l'espace des individus.

En répartissant Λ sur A et B' de manière symétrique, la décomposition est effectuée dans un espace intermédiaire. Considérons l'expression :

$$Y \approx AB'$$
$$y_{ij} \approx a'_i b_j$$

Ce produit scalaire montre que la répartition des projections des points a_i sur les droites b_j approxime la répartition des données initiales pour la variable y_j, quelle que soit la décomposition effectuée.

On peut calculer la qualité d'approximation par le biplot de :

la matrice centrée Y	$\frac{\lambda_1 + \lambda_2}{\sum_{k=1}^{k=r} \lambda_k}$
la matrice de covariance S	$\frac{\lambda_1^2 + \lambda_2^2}{\sum_{k=1}^{k=r} \lambda_k^2}$
la matrice des distances de Mahalanobis pour les individus	$\frac{\lambda_1^0 + \lambda_2^0}{\sum_{k=1}^{k=r} \lambda_k^0} = \frac{2}{r}$

Les propriétés de la méthode GH' (espace des variables)

$$Y \approx AB' = (U)(\Lambda V')$$

Le carré de la norme du vecteur b_j représente n fois la variance de la colonne y_j.

Le cosinus de l'angle entre les vecteurs b_j et b'_j correspond à la corrélation entre les variables y_j et y'_j (comme pour l'ACP, ceci n'est vrai que pour les variables actives; pour les variables supplémentaires, cette propriété n'est vraie que vis-à-vis des axes).

La distance euclidienne entre les points a_i et a'_i est proportionnelle à la distance de Mahalanobis entre les individus y_i et y'_i du tableau de départ.

La distance de Mahalanobis est une distance multidimentionnelle, qui tient compte des corrélations entre les variables ainsi que de leurs variances. Elle tend à transformer un nuage de points de forme allongée en un nuage de forme ronde.

$$\delta^2_{\text{mahalanobis}}(ii') = (y_i - y'_i)' S^{-1} (y_i - y'_i)$$

où S est la matrice de variance-covariance.

Les propriétés de la méthode JK' (espace des individus)

$$Y \approx AB' = (U\Lambda)(V')$$

La distance euclidienne entre les points a_i et a'_i approxime la distance euclidienne entre les individus y_i et y'_i du tableau de départ. Il n'y a pas dans ce cas de propriétés spécifiques aux variables.

Les propriétés de la méthode Symétrique (espace intermédiaire)

$$Y \approx AB' = (U\Lambda^{1/2})(\Lambda^{1/2}V')$$

Cette méthode égalise l'effet des lignes et des colonnes :

Pour chaque axe, les sommes des carrés des écarts à l'axe sont égales pour les individus et les variables. On obtient alors un graphique «équilibré». Hormis la propriété commune à toute les décompositions (les projections des individus sur les variables approximent les répartitions initiales), il n'y pas de propriétés spécifiques à l'interprétation des individus ni des variables.

Lien avec l'ACP

Le biplot est basé sur les mêmes principes de décomposition que l'ACP. On trouve donc des résultats identiques à un coefficient multiplicatif près.

Le biplot fait une recherche des vecteurs et valeurs propres de $Y'Y$, tandis l'ACP fait la recherche sur $Y'DY$, où D est la matrice des poids (en général, les termes de la diagonale valent $1/n$).

	A (individus)	B' (variables)
GH'	U	$\Lambda V'$
Symétrique	$U\Lambda^{1/2}$	$\Lambda^{1/2}V'$
JK'	$U\Lambda$	V'
ACP	$U^*\Lambda^*$	$\Lambda^*V^{*'}$

avec : $Y = U\Lambda V'$ décomposition à partir de $Y'Y$,

$Y = U^*\Lambda^*V^{*'}$ décomposition à partir de $Y'DY$ où $U^* = \sqrt{n}U$, $\Lambda^* = \Lambda/\sqrt{n}$ et $V^* = V$.

On obtient les formules de passage suivantes, dans lesquelles $\text{Fact}_k*(i)$ désigne la coordonnée de l'individu i sur l'axe k et $\text{Fact}_k*(j)$ la covariance (la corrélation en ACP normée) de la variable j avec le facteur k de l'ACP de Y :

a) pour les individus :

$$GH_k(i) = \frac{\text{Fact}_k^*(i)}{\sqrt{n\lambda_k^*}}$$

$$JK_k(i) = \text{Fact}_k^*(i)$$

$$SY_k(i) = \frac{\text{Fact}_k^*(i)}{\sqrt{\sqrt{n\lambda_k^*}}}$$

b) pour les variables

$$GH_k(j) = \sqrt{n}\mathrm{Fact}_k^*(j)$$

$$JK_k(j) = \frac{\mathrm{Fact}_k^*(j)}{\sqrt{\lambda_k^*}}$$

$$SY_k(j) = \frac{\sqrt{\sqrt{n}\mathrm{Fact}_k^*(j)}}{\sqrt{\sqrt{\lambda_k^*}}}$$

La principale différence entre l'ACP est due à la nature des opérations effectuées :

- l'ACP utilise des projections;
- le BIPLOT utilise des approximations.

Un tracé sur le plan 2-3 d'un Biplot n'a plus les propriétés d'un Biplot car on perd dans ce cas la reconstitution des données initiales. C'est pourquoi le Biplot est généralement limité au plan 1-2. Par contre il est possible de faire une représentation dans l'espace en prenant les 3 premiers axes factoriels. On garde dans ce cas toutes les propriétés du Biplot, appelé dans ce cas Bimodel.