

HENRI CAUSSINUS

**Modèles probabilistes et analyse des données
multidimensionnelles. Quelques réflexions
méthodologiques et applications**

Journal de la société statistique de Paris, tome 134, n° 2 (1993),
p. 15-32

http://www.numdam.org/item?id=JSFS_1993__134_2_15_0

© Société de statistique de Paris, 1993, tous droits réservés.

L'accès aux archives de la revue « Journal de la société statistique de Paris » (<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/>

MODÈLES PROBABILISTES ET ANALYSE DES DONNÉES MULTIDIMENSIONNELLES

Quelques réflexions méthodologiques et applications

Henri CAUSSINUS *

*Laboratoire de Statistique et Probabilités,
URA CNRS 745, Université Paul Sabatier, Toulouse*

Prologue

Un jour, ils s'aperçurent qu'ils faisaient sans doute fausse route. Beaucoup alors préconisèrent de changer d'embarcation (on en profiterait pour remplacer l'ancien pavillon, un S d'or sur champ de sable, par un nouveau, aux A et D d'argent sur champ d'azur). D'autres arguèrent qu'il fallait d'abord et surtout changer de cap (ce que certains des précédents avaient d'ailleurs, dans leur enthousiasme, oublié), quitte à radouber convenablement le navire si nécessaire. Un petit nombre voulut carrément couler le bateau (c'était une époque radicale), mais ils ne sont cités que pour mémoire car ils n'eurent finalement pas gain de cause. D'autres enfin, plus nombreux, souhaitèrent ne rien changer (au moins pour l'instant), ni cap ni navire, trouvant l'environnement agréable et la nouveauté par trop dérangeante ; de plus, ils n'étaient pas si sûrs de faire fausse route... Ils se désintéressèrent donc de la discussion, et nous leur rendrons ici la pareille.

J'étais dans le second groupe, parmi ceux croyant que l'on peut parfois (pas toujours) changer de direction en restant dans le même bateau, à condition de disposer voiles et barre de façon adéquate. Je vais essayer d'expliquer ce point de vue, mais, piètre marin, je prendrai pour cela une situation semblable rencontrée par les statisticiens quand de nouveaux problèmes se sont posés à eux suite à une évolution majeure des moyens de recueil et de traitement des données, évolution qui a considérablement modifié maint objectif, en général dans le sens d'une plus grande ambition. Et cette allégorie aura même l'avantage de nous rapprocher de thèmes plus généralement développés dans ce journal...

* Communication du 20 janvier 1993 faite à l'occasion de la remise du Prix 1992 du Statisticien d'expression française décerné par la S.S.P.

Présentation d'ensemble

Il fallait donc changer de cap, sans rejeter les acquis, mais en adaptant les outils aux circonstances nouvelles. Les idées générales, énoncées d'abord, seront ensuite illustrées sur le cas particulier des *analyses en composantes principales* (ACP) considérées comme des techniques visant à projeter un nuage de points sur des sous-espaces convenables de petite dimension.

Ainsi, cet article sera l'occasion de faire le point sur un certain nombre de nos travaux de la dernière décennie. Ceux-ci sont annoncés dans Caussinus (1984), un cours introductif à l'ACP qui, tout en étant orienté vers les aspects exploratoires (essentiellement descriptifs), cherchait à discuter de l'utilité d'un modèle probabiliste adéquat pour la meilleure compréhension et le meilleur « réglage » de cette technique. Si notre réflexion méthodologique datait de plusieurs années, ce texte en inaugurerait la formalisation, tout en restant de portée théorique et pratique limitée. Plus significatifs sont les articles publiés deux ans plus tard (Caussinus, 1986a, b ; Besse, Caussinus, Ferré et Fine, 1986) qui exposent nos points de vue théoriques et en donnent les premières applications. D'autres applications ont suivi (voir par exemple Caussinus et Ferré, 1989, 1992) et nos points de vue méthodologiques se sont quelque peu infléchis sans toutefois s'en trouver bouleversés (Caussinus et Ruiz-Gazen, 1993).

L'ensemble de cette réflexion a bien sûr été stimulé par différents auteurs. Pour la nécessaire remise en cause méthodologique, il faut citer au moins Tukey (1962) et Mallows and Tukey (1982) de même que Benzécri (1973). Nos points de vue sont proches de ceux de Barra (1985) et de son « modèle virtuel » ou Barra et Becker (1986) pour l'importance accordée au « format des données » et pour le concept d'idéomorphisme. D'autres sources seront indiquées plus loin et, surtout, des bibliographies spécialisées plus complètes seront trouvées dans les divers articles cités et dans l'ouvrage édité par Drosbeke, Fichet et Tassi (1992).

Statistique, modèles et analyse des données

On a beaucoup opposé *Statistique* et *Analyse des données*, bien que ceux qui s'intéressent à l'une ou l'autre de ces activités aient clairement un objectif commun : tirer le meilleur parti de données d'expérience ou d'observation. Mais il y a pour cela toute une hiérarchie de méthodes dépendant de la qualité des données comme des objectifs poursuivis, depuis la prise de décision à partir de données éventuellement peu nombreuses mais résultant d'expériences bien structurées, jusqu'à l'exploration d'une masse importante de données d'observation très peu structurées *a priori*. Dans le premier cas, une (ou plusieurs) population homogène précise est supposée génératrice des données ; et c'est elle qui, au-delà des données elles-mêmes, fait l'objet d'investigations. Ainsi, selon Fisher (1947), « l'étude des populations et celle des variations constituent (...) l'objet propre de la statistique » qui est donc « l'étude de populations ou d'agrégats d'individus, plutôt que l'étude d'individus ». Dans le second cas, il est difficile de définir une population originelle, si bien que chaque individu conserve sa spécificité ; l'analyse ne saurait alors porter que sur les seules données, sans référence à une population, et ses objectifs s'en trouvent donc profon-

dément différents. Ils sont de nature préliminaire, exploratoire, par opposition aux objectifs confirmatoires, et surtout généralement plus spécifiques, de la situation précédente (Il est intéressant de noter que Cox (1977) oppose *exploratoire* à *spécifique* plutôt qu'à *confirmatoire*, ce qui est sans doute plus pertinent à plusieurs égards). Mais il semble clair que toutes les situations intermédiaires peuvent se présenter (Quel est le statut de la régression ?) et qu'il y a une continuité naturelle entre *Statistique* et *Analyse des données*, les deux termes n'ayant d'intérêt que pour mettre l'accent sur telle ou telle priorité dans le cadre général de l'*Analyse Statistique des Données*. A l'intérieur de ce domaine, souligner ce qui rassemble plutôt que ce qui sépare, conduit naturellement à rechercher des bases méthodologiques communes.

L'opposition entre *Statistique* et *Analyse des données* se cristallise souvent autour du rôle et de la nature du modèle dans l'approche des problèmes. C'est donc en premier lieu le concept de *modèle* qui retiendra notre attention. Dans son acception première (la seule mentionnée par Littré à peu de nuances près), le modèle est un « objet d'imitation », modèle du peintre ou de vertu. De ce sens originel, nous retiendrons que le modèle est un *guide*, en fait, un guide plus ou moins strict (toute peinture n'est pas « académique »). En statistique, « le modèle est un système de représentation formelle d'idées ou de connaissances » (Larousse); la difficulté opérationnelle d'une telle définition est clairement qu'il y a loin des unes aux autres, de quelques idées vagues à des connaissances avérées. En fait, la recherche scientifique appuyée sur des données a pour objectif général de passer des premières aux secondes; elle est donc concernée par le modèle dans ses diverses acceptions.

Une analyse statistique « classique » (entendons par là celle qui est enseignée comme telle, et particulièrement la théorie de l'estimation et, davantage encore, celle des tests) est fondée sur un modèle, nécessaire pour la justifier, et concerne même le modèle lui-même (qu'on le teste, qu'on estime certains de ses paramètres, ou qu'on prenne n'importe quel type de décision à son égard). La qualité du modèle, son degré de vraisemblance, est importante, d'une part parce que les données, le plus souvent peu nombreuses, contiennent peu d'information par rapport aux objectifs et que cette carence est en quelque sorte palliée par les hypothèses d'un modèle contraignant, d'autre part parce que les réponses que l'on attend ont un certain aspect définitif (les domaines des essais thérapeutiques ou du contrôle de fabrication fournissent des exemples éloquents). La qualité des résultats est ici très largement conditionnée par celle du modèle. D'ailleurs, une bonne part de la recherche statistique est vouée au problème de l'allègement des hypothèses en vue de rendre le modèle plus fiable (techniques non paramétriques) ou à l'art de tirer de bonnes conclusions d'un modèle douteux (robustesse). En bref, dans un *contexte confirmatoire spécifique*, les modèles sont avant tout considérés comme des modèles de connaissance aussi bien dans leurs fonctions de cadrage de l'analyse (« *secondary aspects of a statistical model* » selon la terminologie de Cox, 1977) que dans leurs fonctions de questionnement (« *primary aspects* » au sens de Cox).

La situation est tout autre en présence d'un grand nombre de données, éventuellement peu structurées et recueillies dans un but peu précis. Le contexte est alors *exploratoire*, c'est-à-dire les questions plus ouvertes, sans référence nécessaire à un

modèle *a priori*. Ainsi est-il naturel de dire que « le modèle doit suivre les données, non l'inverse » (Benzécri, 1973) en entendant par là que la *connaissance* procède des données. Il n'en reste pas moins que l'usage d'une certaine forme de modèle *a priori* garde son intérêt pour au moins deux raisons. D'une part, les données au sens strict (valeurs observées) ne sont pas la seule information disponible : le contexte dans lequel elles sont obtenues, leur nature (ex. : fréquences ou mesures), sont des éléments non négligeables dans le choix d'un mode d'analyse ; on peut essayer d'introduire ces informations dans un modèle qui précise le cadre de l'étude. D'autre part, il est illusoire de croire qu'une analyse exploratoire est entièrement ouverte ; les outils d'analyse orientent vers des classes précises de questions (et de réponses) correspondant en tout état de cause à quelques vues *a priori* du statisticien ou de l'usager, même si ces vues englobent de larges perspectives. Les méthodes d'analyse correspondent donc à des modèles implicites qu'il est plus naturel et raisonnable d'explicitier : cela peut permettre à la fois de mieux régler la méthode et d'en mieux apprécier la portée.

Ainsi nous pensons que, même en analyse exploratoire des données, il est utile d'explicitier un modèle qui constitue le cadre de travail et se trouve chargé de guider l'analyse, un modèle aussi « léger » que possible, mais dont il faut cependant garder à l'esprit le caractère très imparfait d'expression de quelques idées souvent extrêmement vagues. Ce modèle a les caractéristiques suivantes :

1) **Le modèle dépend de l'environnement et de la nature des données, ainsi que des objectifs de l'analyse.** C'est en effet un médiateur qui doit prendre en compte les tenants et aboutissants du problème. Le second point est particulièrement important car un modèle *a priori* trop strict et mal adapté peut faire passer à côté de découvertes importantes en limitant le champ de l'analyse.

2) **Le modèle n'est ni présumé exact ni unique.** Le premier point est banal (Voir la remarque de M. Wilk, cité par Tukey, 1962 : « *The hallmark of good science is that it uses models (...) but never believes them* »), mais il est important de discuter quels effets peut avoir un modèle trop inadéquat. Quant à la multiplicité des modèles envisageables, elle découle aussi bien de la multiplicité d'objectifs possibles que de la variété des points de vue sous lesquels on peut les aborder.

Revenons un peu sur l'inadéquation du modèle. On a vu que l'inadéquation aux objectifs pouvait occulter des aspects importants, c'est-à-dire donner lieu à des FAUX « négatifs » ; les FAUX « positifs » apparaîtront au contraire si des présupposés erronés sur la structure des données (ou plutôt sur le phénomène qu'elles représentent) sont susceptibles de se répercuter sur les résultats de l'analyse. Ces faux positifs étant usuellement considérés comme les plus graves dans la déontologie statistique, on retrouve l'importance d'une modélisation aussi souple que possible des aspects *a priori*. Il faut cependant noter qu'il est en général facile de discerner quels résultats dépendent fortement du modèle et quels résultats en dépendent peu. Pour prendre un exemple banal, une hypothèse de normalité permet d'affirmer que la moyenne empirique est le meilleur estimateur sans biais de l'espérance mathématique ; sans cette hypothèse, cet estimateur est toujours sans biais, mais il n'est plus nécessairement le

meilleur (les objectifs peuvent n'être que partiellement atteints) ; un intervalle de confiance calculé sous hypothèse de normalité pourra avoir un coefficient de sécurité faux si celle-ci n'est pas satisfaite, mais restera convenablement centré, etc. (tout cela évidemment si l'on admet au moins l'existence de l'espérance mathématique pour que les objectifs affichés aient un sens !). Même si tous les exemples ne sont pas aussi simples, particulièrement en analyse multidimensionnelle, nous pensons que le statisticien est suffisamment armé pour avoir conscience des dangers d'un FAUX « positif » ; par contre, la mise en évidence de faits intéressants sera d'autant plus aisée qu'un bon cadre de travail aura soigneusement posé l'analyse adaptée, de sorte qu'un modèle bien choisi peut être garant de découvertes qui auraient échappé à une analyse « aveugle » (aveugle soit parce qu'elle manque de l'éclairage d'un modèle adapté, soit parce qu'elle est éblouie par la trop grande clarté d'un modèle abusif !).

Les derniers commentaires concerneront la nature probabiliste du modèle dans le cas de l'analyse de données multidimensionnelles. Disons d'abord que rejeter le modèle probabiliste parce que les données sont nombreuses, au nom de la loi des grands nombres, est non seulement paradoxal, mais encore très réducteur, car il faut le plus souvent éviter de supposer *a priori* que l'on se trouve en présence d'un échantillon au sens strict (indépendance et même loi). On peut arguer que, si l'on met l'accent sur l'exploration d'un nuage de points afin de rechercher des aspects structurels, ceux-ci ont une nature déterministe qui renvoie au second plan le caractère aléatoire des données. Il est clair cependant que, dans une large mesure, « certaines observations auraient pu se trouver différentes sans que les résultats de l'analyse fussent en être modifiés » (Caussinus, 1986b), et c'est à la prise en compte de cet aspect des choses que visera le modèle probabiliste. Une formulation adaptée sera souvent :

$$\text{Données} = \text{Structure} + \text{Bruit}$$

où la partie structurelle sera modélisée de façon algébrique (ou même éventuellement probabiliste comme on en verra un exemple plus loin), tandis que le bruit sera toujours aléatoire par essence puisqu'il représente la partie des données que l'on renonce à « structurer » (dans une phase précise de l'étude). Soulignons que le signe « plus » ci-dessus doit être pris dans un sens très général et ne correspond pas nécessairement à une addition au sens strict comme, par exemple, dans les modèles linéaires d'analyse de la variance (Voir McCullagh and Nelder, 1989, pour l'exemple des modèles linéaires généralisés où cet aspect est fondamental).

Les points de vue développés ci-dessus seront maintenant illustrés et précisés sur l'exemple de l'Analyse en Composantes Principales, ou plutôt sur le problème plus général de la représentation euclidienne d'un nuage de points dans un espace de dimension réduite. En effet, l'ACP « ordinaire » ne donne de ce problème qu'une solution parmi bien d'autres possibles. Sous le nom de *projections révélatrices* (expression proposée par Y. Escoufier comme traduction libre mais évocatrice de l'anglais *Projection Pursuit*), des solutions de nature très différente ont été proposées, solutions qui prennent mieux en compte l'intérêt concret présenté par tel ou tel type de projection (Voir le chapitre 8 de Droesbeke *et alii*, 1992). Pour leurs possibilités limitées, Sibson (1984) a d'ailleurs fort justement critiqué les méthodes du « second

ordre », c'est-à-dire celles qui, comme l'ACP usuelle, sont exclusivement basées sur l'analyse de matrices de variances et covariances, fonctions du second degré des données. Nous verrons en particulier que la considération d'un modèle simple permet d'introduire naturellement des ACP qui ne sont plus du second ordre, en ce sens qu'elles font appel à des matrices de dispersion qui sont des fonctions plus complexes des données. Un autre modèle servira à évaluer les performances que l'on peut attendre de ces dernières techniques.

Avant de détailler cet exemple, il convient de préciser qu'il est loin de constituer le seul problème d'analyse de données multidimensionnelles susceptible de bénéficier du support de modèles probabilistes : pour une présentation des questions de classification dans un esprit analogue, on peut se référer aux chapitres 6 et 7, par G. Celeux, de l'ouvrage édité par Dreesbeke *et alii* (1992).

Modèles pour l'Analyse en Composantes Principales

Les données sont n vecteurs X_i ($i = 1, \dots, n$) de \mathbb{R}^p identifiés aux matrices colonnes de leurs coordonnées dans une base fixée ; les coordonnées de X_i sont les p mesures effectuées sur l'unité statistique (ou individu) numéro i . On notera respectivement \bar{X} et V le vecteur moyenne empirique et la matrice des variances et covariances empiriques, c'est-à-dire :

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, \quad V = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})'$$

Dans la seconde formule ci-dessus, comme par la suite, la transposée d'une matrice A est notée A' . Par ailleurs, si M est une matrice $p \times p$ symétrique définie positive et x un vecteur $p \times 1$, on notera $\|x\|_M^2 = x'Mx$ le carré de la norme euclidienne de x au sens de la métrique M .

On rappelle que, si l'espace des individus \mathbb{R}^p est muni de la métrique M , l'ACP centrée d'ordre q ($q < p$) du nuage des X_i consiste à :

- déterminer le sous-espace engendré par q vecteurs propres de VM associés aux q plus grandes valeurs propres (sous-espace principal d'ordre q),
- projeter M -orthogonalement sur ce sous-espace les vecteurs $X_i - \bar{X}$ afin d'obtenir une représentation euclidienne q -dimensionnelle des individus.

Cette présentation de l'ACP ne concerne que l'aspect « individus » auquel on s'intéresse essentiellement ici, l'aspect « variables » et les questions de dualité restant secondaires dans cet article. Dans une optique exploratoire, on se donne en effet comme objectif principal de visualiser les n individus dans un espace de dimension q inférieure à p . A partir de cette représentation, on cherche à appréhender au mieux la « structure » du nuage des X_i ; comment se situent les individus les uns par rapport aux autres, quels sont les individus extrêmes, y a-t-il des regroupements, des individus atypiques (*outliers*) ?, sont autant de questions auxquelles on essaie de répondre, parmi bien d'autres. L'argument le plus classique à la base de l'ACP est de nature non probabiliste : celle-ci, telle qu'elle est décrite plus haut, projette le nuage des

MODÈLES PROBABILISTES ET ANALYSE DES DONNÉES

individus (centrés) sur un sous-espace de dimension q , sélectionné pour déformer le moins possible ce nuage selon un critère d'inertie. Si $\Pi_{E_q}^M$ désigne le projecteur M -orthogonal sur E_q , on rend minimum

$$\sum_{i=1}^n \|X_i - \bar{X} - \Pi_{E_q}^M(X_i - \bar{X})\|_M^2$$

pour E_q parcourant l'ensemble des sous-espaces de \mathbb{R}^p de dimension q .

Mais ce point de vue simple fournit difficilement des arguments pour régler convenablement l'analyse. Par exemple, il est impuissant à lui seul à suggérer une métrique M convenable, problème majeur qui a suscité bien des travaux (Parmi ceux-ci, citons les contributions de P. Besse, F. Houllier, J.B. Kazmierczak, M. Maurin, R. Sabatier et S. Winsberg, dans le numéro spécial de *Statistique et Analyse des données* de décembre 1987, préparé par Y. Escoufier). De même, il ne permet pas de choisir q de façon objective ; le très grand nombre de « méthodes » empiriques proposées pour cela (Voir Ferré, 1988, pour une revue de celles-ci) nous montre bien l'embaras des praticiens.

De façon générale, il n'est pas sûr que l'ACP fournisse la vue la plus appropriée pour mettre en évidence une structure particulière, même très forte, des données. Les figures 1 et 2 en témoignent sur des exemples schématiques (Il faut voir ces figures comme si chacun des deux axes correspondait à un sous-espace de dimension au moins égale à deux, et non *un*) : l'ACP « usuelle » donne une projection sur l'axe (sous-espace) horizontal qui préserve le maximum d'inertie mais ne fait apparaître ni les deux groupes, pourtant assez nettement séparés, de la figure 2, ni les deux individus les plus atypiques de la figure 1.

Figure 1

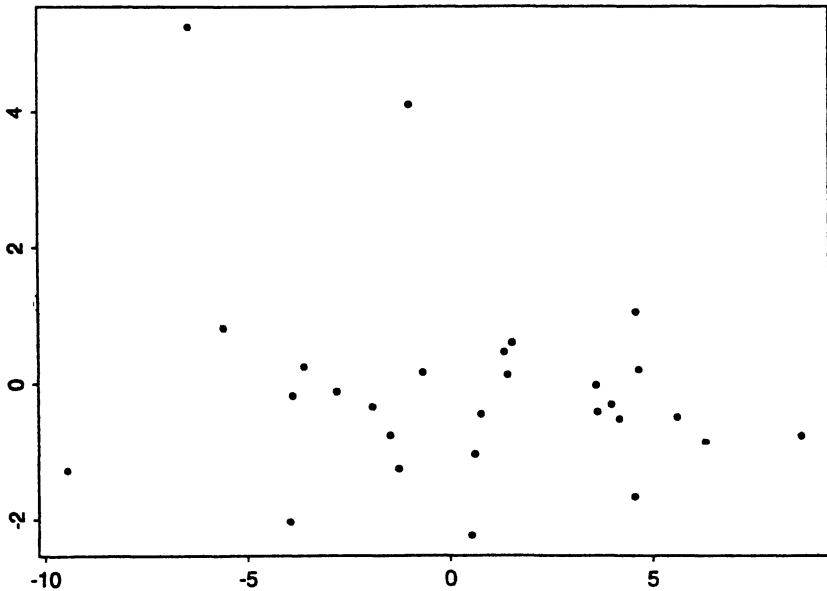
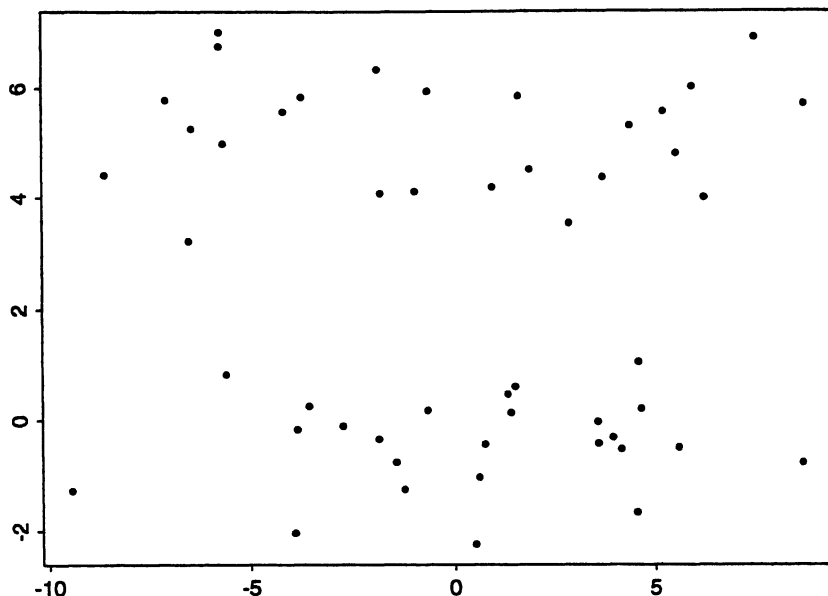


Figure 2



On verra plus loin qu'une ACP avec une métrique M convenable peut faire ressortir de telles structures *quand elles ne sont pas connues a priori*. Remarquons en passant que, si dans le cas de la figure 2, il est connu que les points sont divisés en deux groupes et quels sont ces groupes, l'Analyse Factorielle Discriminante en fournit une projection qui les met en évidence, mais celle-ci a peu d'intérêt *exploratoire* puisqu'elle n'exhibe qu'un fait connu ; elle peut, par contre, avoir un intérêt *confirmatoire* en visualisant la plus ou moins bonne séparation des deux groupes et complétant ainsi, par exemple, un test de Hotelling.

Revenons maintenant au problème exploratoire et essayons d'écrire un modèle qui justifie et qui guide les préoccupations exprimées plus haut.

Modèle 1

Les X_i sont des vecteurs aléatoires indépendants tels que :

(a) il existe une variété affine F_q de \mathbb{R}^p , de dimension q ($q < p$), telle que $\mathbf{E}(X_i) \in F_q$ pour tout $i = 1, \dots, n$;

(b) $\text{Var}(X_i) = \sigma^2 W$, où W est une matrice symétrique définie positive et $\sigma > 0$.

Dans ce modèle, F_q , σ et les n moyennes $x_i = \mathbf{E}(X_i)$ sont des paramètres inconnus, tandis que q et W sont supposés connus. Cette dernière condition est en pratique gênante et sera discutée plus loin. L'hypothèse (b) peut être remplacée sans difficulté par :

MODÈLES PROBABILISTES ET ANALYSE DES DONNÉES

(b') $\text{Var}(X_i) = \frac{\sigma^2}{w_i} W$ où les w_i sont des poids strictement positifs donnés.

Le Modèle 1 écrit simplement que l'observation X_i est aléatoire autour d'une moyenne inconnue x_i qui est *caractéristique de l'individu i* , la différence $X_i - x_i$ étant un « bruit », vecteur aléatoire de moyenne nulle et variance $\sigma^2 W$; la « forme » du bruit est donc donnée au second ordre par W et son amplitude par σ . Un des premiers à avoir souligné le réalisme de ce modèle est sans doute Young (1940).

Le Modèle 1 peut être considéré comme implicite lorsqu'une ACP est pratiquée afin de visualiser les n individus dans un espace à q dimensions. En effet, se satisfaire d'une telle représentation, c'est admettre qu'elle fournit une image des aspects pertinents des individus, tandis que les dimensions délaissées ne contiennent que du bruit. Il est facile de montrer que l'ACP centrée des X_i , avec une métrique M sur \mathbb{R}^p , fournit l'estimation des moindres carrés (pour la métrique M) des paramètres F_q et x_i ($i = 1, \dots, n$) : l'estimation de F_q est \hat{F}_q passant par \bar{X} et parallèle au sous-espace principal de dimension q (disons \hat{E}_q) engendré par q vecteurs propres de VM associés aux q plus grandes valeurs propres ; l'estimation \hat{x}_i de x_i est la projection M -orthogonale de X_i sur \hat{F}_q , soit $\hat{x}_i - \bar{X} = \Pi_{\hat{E}_q}^M(X_i - \bar{X})$: c'est la représentation de l'individu i dans cette ACP.

Jusqu'ici, le Modèle 1 est seulement une façon parmi d'autres de concevoir l'ACP ; mais la considération de ce modèle permet d'apporter quelques éléments nouveaux par rapport au point de vue purement descriptif. Dans le cadre du Modèle 1, il est clair que l'analyse statistique substitue le vecteur \hat{x}_i au vecteur x_i caractéristique de l'individu i ; il est alors naturel de mesurer la qualité globale de la représentation par l'erreur quadratique moyenne

$$f_q(M, A) = \sum_{i=1}^n \mathbf{E} (\| \hat{x}_i - x_i \|_A^2)$$

où A est une métrique quelconque sur \mathbb{R}^p .

Cette quantité dépend de M par \hat{x}_i , projection M -orthogonale de X_i sur \hat{F}_q qui dépend aussi de M . Le choix optimal de M est celui qui rend $f_q(M, A)$ minimum. Il est difficile de traiter la recherche d'un tel M dans toute sa généralité, mais on peut donner une solution approchée pour σ petit en minimisant la partie principale de $f_q(M, A)$. On montre alors que, indépendamment de A , la métrique optimale est à un coefficient multiplicatif près (Besse, Caussinus, Ferré, Fine, 1987, 1988) : $M = W^{-1}$ (propriété analogue à celle de Gauss-Markov pour le modèle linéaire).

Le même résultat est obtenu par Fine et Pousse (1992) pour une asymptotique en n ($n \rightarrow +\infty$, σ fixé). Ils prouvent même que le choix de métrique ci-dessus est le seul pour lequel \hat{F}_q est un estimateur consistant (presque sûrement convergent) de F_q .

Dans la pratique, le résultat précédent n'est utilisable que si W est connu ou peut être approché de façon convenable, par exemple grâce à quelques informations sur

la nature des données. On peut ainsi trouver de nouvelles justifications aux métriques utilisées en Analyse Discriminante ou Analyse des Correspondances (cf. Caussinus, 1986a, b), ou trouver de nouvelles bases à l'Analyse des Correspondances Sphériques de Domenges et Volle (1979), ou encore à l'ACP des paramètres d'un modèle (Caussinus et Ferré, 1989, 1992) au-delà des justifications plus simplement descriptives de Houllier (1987). Parmi d'autres applications possibles, citons encore l'approche du modèle d'association de Goodman proposée par Baccini et Khoudraji (1992). Néanmoins, on peut dire que, jusqu'ici, l'utilisation du Modèle 1 apporte peu d'éléments strictement nouveaux puisqu'il a surtout servi à reformuler des techniques connues par ailleurs. Mais il permet aussi de déboucher sur des développements plus originaux que nous indiquons maintenant.

Notons d'abord que l'explicitation du Modèle 1 montre clairement certaines limites de l'ACP. Par exemple, cette technique ne se trouve pleinement justifiée que si les bruits sont de même variance $\sigma^2 W$ pour chaque individu. Des variances de la forme $\frac{\sigma^2}{w_i} W$ conduiront à une ACP pondérée, susceptible aussi de fournir une approche raisonnable dans certains cas plus généraux (Caussinus et Ferré, 1989, 1992). Pour certaines situations où l'hypothèse (a) du Modèle 1 subsiste mais non l'hypothèse (b), ni (b'), on peut remplacer l'algorithmique usuelle (estimation de moindres carrés) par un autre type d'estimation (Voir, par exemple, Gilula et Haberman, 1986, pour l'analyse des correspondances d'un tableau de fréquences).

Une application importante de la mise en évidence du Modèle 1 nous semble concerner le choix de la dimension de représentation q optimale. Toujours dans l'optique de l'optimisation de la représentation des x_i par les \hat{x}_i , ayant choisi $M = W^{-1}$, et prenant par exemple $A = W^{-1}$ pour simplifier, on peut chercher q qui minimise $f_q(W^{-1}, W^{-1})$, ou plutôt une estimation convenable de cette quantité. Le problème est abordé dans Ferré (1989) et certaines applications apparaissent dans Caussinus et Ferré (1989, 1992). Le fait important au plan méthodologique est que, malgré la plus grande fidélité de la représentation par rapport aux données, ce n'est pas nécessairement avec $q = p$ qu'on obtient le graphique le plus pertinent, si bien que la réduction de dimension par l'ACP peut être utile non seulement pour *simplifier* la représentation mais encore pour l'*améliorer*. Cela provient d'une prise en compte convenable du bruit ; malheureusement celle-ci n'est réalisable que dans certaines situations particulières (mais réalistes) où $\sigma^2 W$ est connu ou estimable.

Dans tous les exemples de choix de la métrique M rencontrés plus haut, ce sont avant tout des considérations sur la nature *objective* des données qui ont conduit vers une sélection appropriée (hypothèses de distribution multinomiale de fréquences pour les Analyses de Correspondances, connaissance de groupes homogènes pour l'Analyse Discriminante, etc.). Mais on peut aussi choisir M pour régler l'analyse afin de faire ressortir une *éventuelle* structure des x_i , *si elle existe*. Le choix de la métrique se trouve alors guidé en premier lieu par l'objectif visé.

De tels choix de métrique ont d'abord été proposés de façon empirique pour traiter les situations schématisées respectivement par les figures 1 et 2 ci-dessus.

MODÈLES PROBABILISTES ET ANALYSE DES DONNÉES

(1) Yenyukov (1988) propose de chercher à faire ressortir les valeurs atypiques en prenant pour M l'inverse d'une estimation robuste de la matrice de dispersion, ce qui revient bien à chercher une estimation de W (à un facteur près) quand tous les x_i sont égaux sauf peut-être quelques-uns d'entre eux correspondant à des individus atypiques.

(2) Art, Gnanadesikan et Kettenring (1982) proposent une estimation de W lorsque les x_i sont supposés égaux à l'intérieur de groupes homogènes sans que les regroupements d'individus ni même le nombre de groupes soient connus. Prendre pour M l'inverse de cette estimation revient à faire de l'analyse factorielle discriminante malgré la méconnaissance des groupements éventuels, ce qui donne sa pleine valeur exploratoire à la méthode ainsi chargée de révéler les groupes, et les situer les uns par rapport aux autres, s'ils existent.

Les techniques ci-dessus ressortissent donc au domaine des *projections révélatrices* (voir, par exemple, Dreesbeke et al., 1992, chap. 8). Nous les avons quelque peu simplifiées, afin surtout d'en permettre un usage routinier, en introduisant les métriques données par les inverses des matrices suivantes (Caussinus et Ruiz, 1990) :

$$S(\beta) = \frac{\sum_{i=1}^n K(\|X_i - \bar{X}\|_{\hat{V}^{-1}}) (X_i - \bar{X})(X_i - \bar{X})'}{\sum_{i=1}^n K(\|X_i - \bar{X}\|_{\hat{V}^{-1}})} \quad (1)$$

$$T(\beta) = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^n K(\|X_i - X_j\|_{\hat{V}^{-1}}) (X_i - X_j)(X_i - X_j)'}{\sum_{i=1}^{n-1} \sum_{j=i+1}^n K(\|X_i - X_j\|_{\hat{V}^{-1}})} \quad (2)$$

où K est une fonction décroissante de \mathbb{R}_+ dans lui-même et β un paramètre réel fixé par l'analyste (l'étude mathématique fournira quelques indications sur les valeurs utiles de β). Nous avons été amenés à privilégier la fonction K définie par

$$K(x) = \exp\left(-\frac{\beta x}{2}\right),$$

seule considérée plus bas.

Cependant, si le Modèle 1 a permis d'introduire de telles métriques, il s'avère inapte à faire clairement ressortir les propriétés des ACP correspondantes. Pour cela, nous allons nous appuyer sur un second modèle.

Modèle 2

Les X_i sont n vecteurs aléatoires indépendants de même loi $\int N_p(x, W) dP(x)$ où la probabilité P est concentrée sur une variété affine F_q de dimension $q < p$.

MODÈLES PROBABILISTES ET ANALYSE DES DONNÉES

La loi de X_i apparaît donc comme un mélange de lois normales de matrice de variances-covariances W , et de moyenne aléatoire de distribution inconnue P concentrée sur la variété affine inconnue F_q . Concrètement, on peut se représenter x_i comme tiré au hasard dans F_q (la loi P étant *a priori* quelconque, cette hypothèse est en pratique équivalente à l'hypothèse (a) du Modèle 1), X_i étant ensuite aléatoire, normale de moyenne x_i et dispersion W . Dans ce dernier modèle, la loi P modélise la *structure* de l'ensemble des individus. L'objectif sera donc de représenter ces individus dans un espace de dimension q , estimant F_q , afin de visualiser au mieux la distribution P . Conditionnellement à x_i , la loi $N_p(0, W)$ de $X_i - x_i$ exprime le bruit. Ici, W n'est pas supposé connu, mais la distribution du bruit est normale. Cela signifie que, conformément au principe des projections révélatrices (voir Huber, 1985 ; Friedman, 1987 ; Jones and Sibson, 1987), nous appelons *structure* les aspects « non normaux » (non gaussiens) des données et c'est cela que nous cherchons à représenter. Dans le cadre du Modèle 2, il est possible de montrer que l'ACP de métrique $S_n^{-1}(\beta)$ ou $T_n^{-1}(\beta)$ y parvient dans certaines situations. De façon précise, on étudie ce qui se passe lorsque n tend vers l'infini et l'on cherche sous quelles hypothèses l'analyse proposée *fonctionne*, c'est-à-dire conduit à projeter les données sur une variété affine qui converge vers F_q , et à les projeter au sens d'une métrique qui permet de diminuer au maximum l'effet du bruit. L'étude fournit en outre des indications sur les valeurs souhaitables du paramètre β . Quelques théorèmes en ce sens sont donnés dans Caussinus et Ruiz (1993). A titre d'exemple, en voici deux.

- Une propriété de l'ACP de métrique $S^{-1}(\beta)$

Dans le cadre du Modèle 2, s'il existe un vecteur x_0 de F_q tel que $P(\{x_0\}) = p$, alors il existe un intervalle $]p_0, 1[$ et un intervalle $]0, \beta_0[$ tels que l'ACP de métrique $S^{-1}(\beta)$ fonctionne si $p \in]p_0, 1[$ et $\beta \in]0, \beta_0[$.

En pratique, cela signifie qu'une proportion p assez grande d'individus ($p > p_0$) est de moyenne x_0 , tandis que les autres (individus dits atypiques en proportion $1 - p$) ont une moyenne différente répartie selon la loi P dans F_q (à part son atome en x_0 , P est quelconque, discrète ou continue). Alors, pour β petit, l'ACP d'ordre q de métrique $S^{-1}(\beta)$ fournit une estimation convenable de F_q pour n assez grand et permet de visualiser au mieux les individus atypiques. L'exemple 1 ci-dessous fournit une illustration de cette propriété. Notons que, dans ce type d'analyse, β doit donc être assez petit (mais non nul évidemment).

- Une propriété de l'ACP de métrique $T^{-1}(\beta)$

Dans le cadre du Modèle 2, notons Q la loi de la différence de deux vecteurs aléatoires indépendants et chacun de loi P . Si, en projection sur F_q , la variance de P est assez grande par rapport à W , et $Q(\{0\}) \neq 0$, alors il existe $\beta_0 > 0$ tel que l'ACP de métrique $T^{-1}(\beta)$ fonctionne pour $\beta > \beta_0$.

En pratique, cela implique d'abord que β doit être ici assez grand (β voisin de 2 semble le plus efficace). Par ailleurs, la condition sur les variances signifie que le

bruit ne vient pas *noyer* la partie structurale. En fait, s'il en était ainsi, il est vraisemblable qu'il serait impossible de mettre en évidence à l'œil cette dernière, même en projetant les données sur le *vrai* F_q . Finalement, la condition $Q(\{0\}) \neq 0$ est réalisée par exemple si P est discrète, c'est-à-dire si les données sont constituées de groupes homogènes et c'est pour détecter une telle structure que la méthode a été conçue. Mais $Q(\{0\}) \neq 0$ est réalisée pour d'autres lois P , en fait dès que P possède un atome. Enfin, cette condition suffisante n'est pas nécessaire et l'ACP fonctionne encore pour d'autres lois P , par exemple dans certains cas où P est concentrée sur une courbe de F_q . L'exemple 2 ci-dessous illustre cette dernière situation, puisque P est concentrée sur une parabole dans un espace à 2 dimensions. Il montre que la méthode est susceptible de mettre en évidence certaines relations non linéaires, une question dont l'intérêt est mentionné par D.R. Cox (discussion de Huber, 1985) mais qui a été peu explorée jusqu'ici.

Au cours de la démonstration des propriétés ci-dessus, on voit que les valeurs propres correspondant au sous-espace qui permet d'estimer F_q sont supérieures à $1 + \beta$ (resp. $\frac{1}{2} + \beta$) pour l'ACP avec métrique $S^{-1}(\beta)$ (resp. $T_n^{-1}(\beta)$). Cela suggère que les valeurs propres du même ordre de grandeur que $1 + \beta$ (resp. $\frac{1}{2} + \beta$), ou *a fortiori* plus petites, sont vraisemblablement associées à du bruit, et fournit donc une technique (heuristique pour le moment) de choix de la dimension adéquate q .

Exemple 1

On a simulé 40 vecteurs de \mathbb{R}^8 de même loi normale et 3 vecteurs supplémentaires de moyennes différentes. Les variances et covariances sont telles que la matrice de dispersion des 43 vecteurs est proche de l'identité si bien que, comme dans la figure 1, l'ACP ordinaire (ou même réduite) est inapte à faire ressortir les individus atypiques qui sont les numéros 41, 42 et 43 (figure 3). Par contre, l'ACP de métrique $S^{-1}(0,1)$ les fait parfaitement apparaître (figure 4). De plus, la suite des valeurs propres est : 3,21 1,71 1,47 1,16 1,08 1,03 1,01 0,95. En comparant à $\beta + 1 = 1,1$, elle suggère que les dimensions au-delà de la troisième ne contiennent que du bruit, que les deux premières dimensions contiennent une structure marquée, tandis que la troisième peut ou non être utile (elle précise seulement un peu les positions respectives des individus 41 et 42 et n'est pas donnée ici).

Exemple 2

On a simulé $n = 100$ vecteurs indépendants de \mathbb{R}^5 de loi $N_5(0, I_5)$ puis, pour chaque $i = 1, \dots, 100$, aux 5 coordonnées X_i^j ($j = 1, \dots, 5$), on a adjoint une sixième

coordonnée X_i^6 telle que
$$X_i^6 = \frac{1}{5\sqrt{2}} \left(\sum_{j=1}^5 X_i^j \right)^2.$$

MODÈLES PROBABILISTES ET ANALYSE DES DONNÉES

Figure 3. Exemple 1. ACP ordinaire : premier plan principal

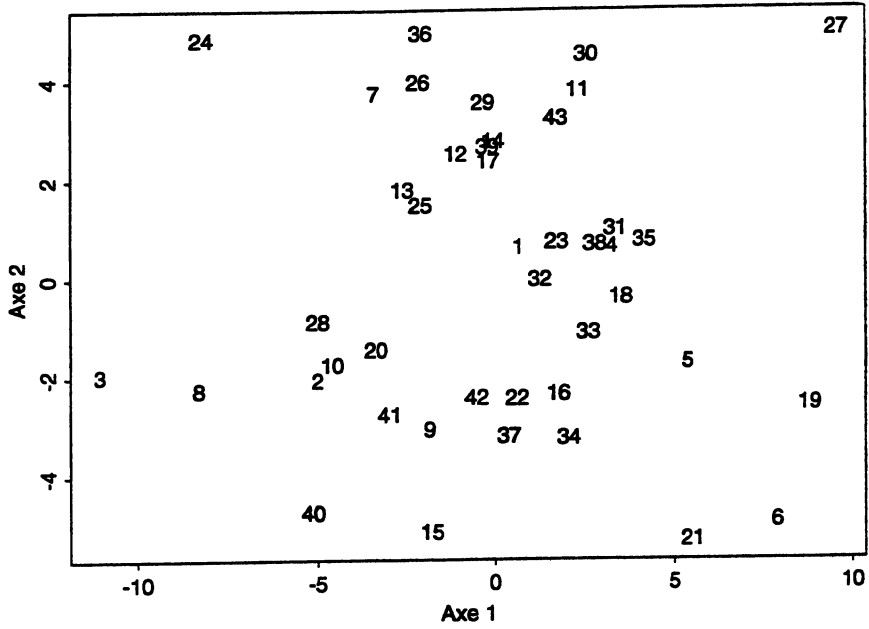
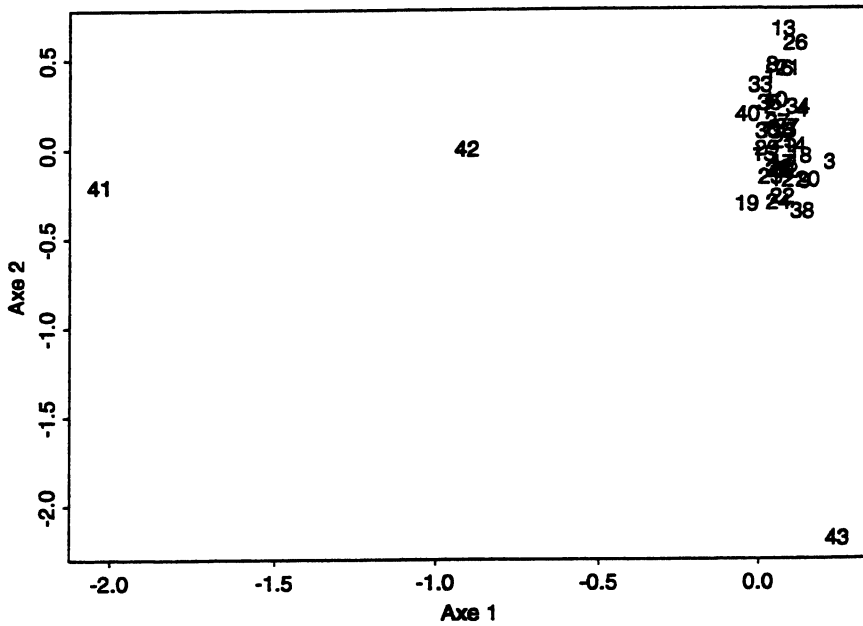


Figure 4. Exemple 1. ACP de métrique $S^{-1}(0, 1)$: premier plan principal



MODÈLES PROBABILISTES ET ANALYSE DES DONNÉES

On a ainsi obtenu 100 vecteurs aléatoires X_i de \mathbb{R}^6 , dont il est facile de vérifier que la matrice des variances et covariances est I_6 , l'identité d'ordre 6. Il n'est donc pas étonnant que l'ACP usuelle ne fasse ressortir aucune structure précise (*figure 5*), tandis que le premier plan principal de l'ACP avec métrique T^{-1} (2) met clairement en évidence une relation parabolique entre les deux coordonnées (*figure 6*). Dans cette dernière analyse, la première coordonnée est une combinaison linéaire des six variables de coefficients

$$(-0,02 \ 0,01 \ 0,09 \ 0,03 \ -0,03 \ 0,99),$$

tandis que la seconde en est une combinaison linéaire de coefficients

$$(0,42 \ 0,51 \ 0,50 \ 0,42 \ 0,34 \ -0,15);$$

la représentation bidimensionnelle associée fait donc apparaître approximativement la relation liant la sixième coordonnée aux cinq premières. Par ailleurs, $\beta + \frac{1}{2} = 2,5$ et la suite des valeurs propres de la dernière ACP (9,13 3,87 2,99 2,62 2,35 2,24) suggère que les dimensions d'ordre supérieur à deux contiennent essentiellement du bruit, ce qui est encore cohérent avec la nature des données simulées.

Figure 5. Exemple 2. ACP ordinaire : premier plan principal

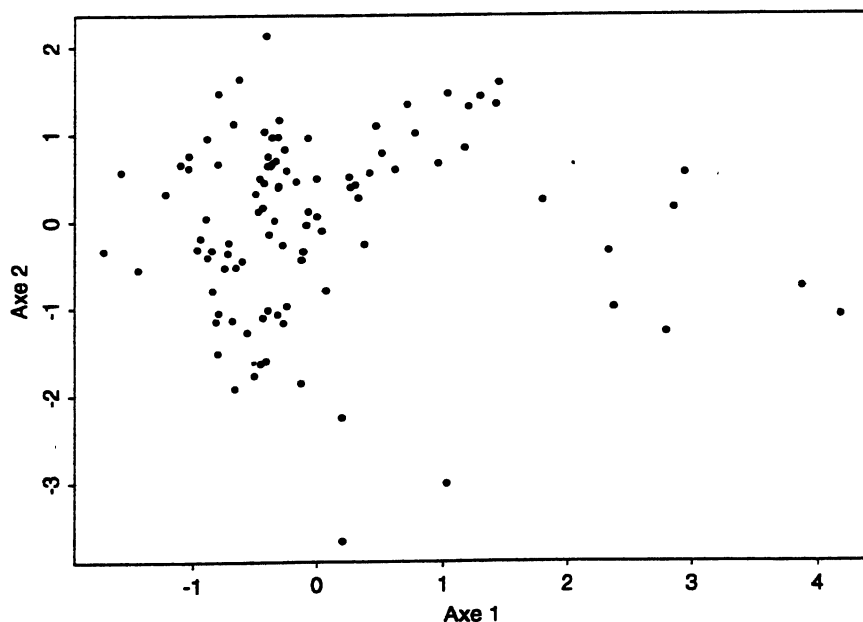
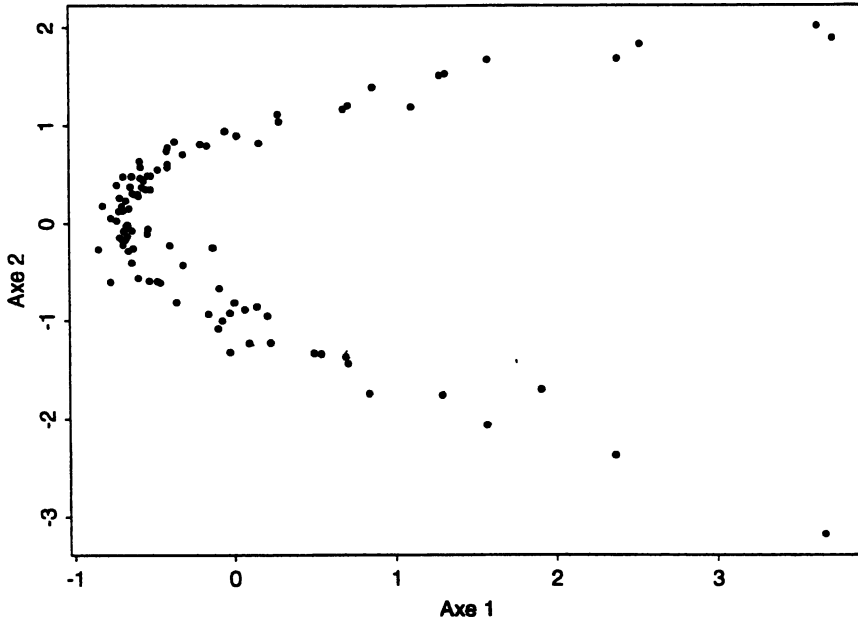


Figure 6. Exemple 2. ACP de métrique $T^{-1}(2)$: premier plan principal



RÉFÉRENCES

- ART D., GNANADESIKAN R., KETTENRING J.R. (1982) "Data-based Metrics for Cluster Analysis", *Utilitas Mathematica*, 21, A, 75-99.
- BACCINI A. and KHOUDRAJI A. (1992) "A Least Squares Procedure for Estimating the Parameters in the RC-Association Model for Contingency Tables", *Computational Statistics*, 287-300.
- BARRA J.-R. (1985) "Méthodes statistiques en psychiatrie - Modèles virtuels", in *Model Choice*, Proceedings of the 4th Franco-Belgian Meeting of Statisticians, Publications des Facultés universitaires Saint-Louis, Bruxelles.
- BARRA J.-R. et BECKER M. (1986) "Conception d'un logiciel d'assistance intelligente en analyse factorielle et canonique", *Data Analysis and Informatics*, IV, 721-732, Diday et al. (Éds.), North-Holland, Amsterdam.

MODÈLES PROBABILISTES ET ANALYSE DES DONNÉES

- BENZECRI J.-P. (1973) *L'analyse des données, 2 : l'Analyse des Correspondances*, Dunod, Paris.
- BESSE Ph., CAUSSINUS H., FERRE L. et FINE J. (1986) "Some Guidelines for Principal Component Analysis", *COMPSTAT 86*, de Antoni *et al.* (Éds), 23-30, Physica-Verlag, Heidelberg.
- BESSE Ph., CAUSSINUS H., FERRE L. et FINE J. (1987) "Sur l'utilisation optimale de l'Analyse en Composantes Principales", *C.R. Acad. Sc. Paris*, 459-562.
- BESSE Ph., CAUSSINUS H., FERRE L. et FINE J. (1988) "Principal Components Analysis and Optimization of Graphical Displays", *Statistics*, 301-312.
- CAUSSINUS H. (1984) "Analyses en composantes principales ; quelques réflexions sur la part des modèles probabilistes", *Publications du Laboratoire de Statistique et Probabilités*, 01-84, Toulouse.
- CAUSSINUS H. (1986a) "Models and Uses of Principal Component Analysis" (with discussion), *Multidimensional Data Analysis*, J. de Leeuw *et al.* (Éds), 149-178, DSWO, Press, Leiden.
- CAUSSINUS H. (1986b) "Quelques réflexions sur la part des modèles probabilistes en analyse des données", *Data Analysis and Informatics*, IV, 151-165, Diday *et al.* (Éds), North-Holland, Amsterdam.
- CAUSSINUS H. et FERRE L. (1989) "Analyse en Composantes Principales d'individus définis par les paramètres d'un modèle", *Statistique et Analyse des Données*, 19-28.
- CAUSSINUS H. and FERRE L. (1992) "Comparing the Parameters of a Model for Several Units by Means of Principal Components Analysis", *Computational Statistics & Data Analysis*, 269-280.
- CAUSSINUS H. and RUIZ A. (1990) "Interesting Projections of Multidimensional Data by Means of Generalized Principal Component Analysis", *COMPSTAT 90*, 121-136, Physica-Verlag, Heidelberg.
- CAUSSINUS H. and RUIZ-GAZEN A. (1993) "Metrics for Finding Typical Structures by Means of Principal Component Analysis", *à paraître*.
- COX D.R. (1977) "The Role of Significance Tests", *Scand. J. Statist.*, 49-70.
- DOMENGES D. et VOLLE M. (1979) "Analyse factorielle sphérique, une exploration", *Annales de l'INSEE*, 3-84.
- DROESBEKE J.-J., FICHET B. et TASSI Ph. (1992) "Modèles pour l'Analyse des Données Multidimensionnelles", *Economica*, Paris.
- FERRE L. (1988) "Le point sur les méthodes de choix de dimension en ACP", *Publication du Laboratoire de Statistique et Probabilités*, n° 01-88, Université Paul Sabatier, Toulouse.
- FERRE L. (1989) "Choix de la dimension optimale pour certains types d'analyses en composantes principales", *C. R. Acad. Sc. Paris*, 959-964.
- FISHER R.A. (1947) *Les méthodes statistiques adaptées à la recherche scientifique*, Presses Universitaires de France.
- FINE J. and POUSSE A. (1992) "Asymptotic Study for the Multivariate Functional Model. Application to the Metric Choice in Principal Component Analysis", *Statistics*, 63-83.
- FRIEDMAN J.H. (1987) "Exploratory Projection Pursuit", *Journal of the American Statistical Association*, 249-266.
- GILULA Z. and HABERMAN S.J. (1986) "Canonical Analysis of Contingency Tables by Maximum Likelihood", *Journal of the American Statistical Association*, 780-788.

MODÈLES PROBABILISTES ET ANALYSE DES DONNÉES

- HOULLIER F. (1987) "Comparaison de courbes et de modèles de croissance ; choix d'une distance entre individus", *Statistique et Analyse des Données*, 17-36.
- HUBER P.J. (1985) "Projection Pursuit (with discussion)", *Ann. Statist.*, 435-525.
- JONES M.C. and SIBSON R. (1987) "What is Projection Pursuit ? (with discussion)", *J. R. Statist. Soc., A*, 1-36.
- Mc CULLAGH M. and NELDER J.A. (1989) *Generalized Linear Models*, Chapman and Hall, London.
- MALLOWS C.L. and TUKEY J.W. (1982) "An Overview of Techniques of Data Analysis, Emphasizing its Exploratory Aspects", *Some Recent Advances in Statistics*, J. Tiago de Oliveira *and al.* (Éd.), 111-172, Academic Press London.
- SIBSON R. (1984) "Present Position and Potential Developments : Some Personal Views - Multivariate Analysis", *J.R. Statist. Soc., A*, 198-207.
- TUKEY J.W. (1962) "The Future of Data Analysis", *Ann. Math. Stat.*, 1-67.
- YENYUKOV I.S. (1988) "Detecting Structures by means of Projection Pursuit", *COMPSTAT 88*, Physica-Verlag, Heidelberg, 47-58.
- YOUNG G. (1940) "Maximum Likelihood Estimation and Factor Analysis", *Psychometrika*, 49-53.