REVUE DE STATISTIQUE APPLIQUÉE

H. CAUSSINUS
S. HAKAM
A. RUIZ-GAZEN

Projections révélatrices contrôlées recherche d'individus atypiques

Revue de statistique appliquée, tome 50, n° 4 (2002), p. 81-94
<http://www.numdam.org/item?id=RSA_2002__50_4_81_0>

Rev. Statistique Appliquée, 2002, **L** (4), 81-94

PROJECTIONS RÉVÉLATRICES CONTRÔLÉES RECHERCHE D'INDIVIDUS ATYPIQUES

H. CAUSSINUS[1], S. HAKAM[1,2], A. RUIZ-GAZEN[1,3]

[1] *Laboratoire de Statistique et Probabilités, U.M.R. - C.N.R.S. C5583, Université Paul Sabatier, 118, route de Narbonne, 31062 Toulouse cedex 4*
[2] *Département de Mathématique et Informatique, Faculté des Sciences, B.P. 1014, Ave. Ibn Battouta, Rabat, Maroc*
[3] *Gremaq, U.M.R. - C.N.R.S. C5604, Université Toulouse I, 21, allée de Brienne, 31000 Toulouse*

RÉSUMÉ

En présence d'un tableau individus × variables, les méthodes de projections révélatrices recherchent des projections exhibant au mieux d'éventuelles structures particulières de la distribution des individus. Caussinus et Ruiz-Gazen ont proposé à cet effet des analyses en composantes principales généralisées par un choix convenable de métriques. Dans cet article, nous rappelons la technique qu'ils préconisent pour la détection graphique et le positionnement d'individus (ou de petits groupes d'individus) atypiques. Nous montrons ensuite comment l'affiner à partir de résultats théoriques récents d'Hakam (2002) : ceux-ci permettent de contrôler les projections significatives, c'est-à-dire la dimension à retenir et les décisions susceptibles d'être tirées des graphiques obtenus. Des exemples (simulé et réel) illustrent la méthode et aident à en discuter la portée.

Mots-clés : *Analyse en composantes principales, Projections révélatrices, Valeurs aberrantes ou atypiques, Techniques graphiques.*

ABSTRACT

Let us consider an individual × variable array. Projection Pursuit aims to find low-dimensional projections displaying interesting features in the structure of the units distribution. For that, Caussinus and Ruiz-Gazen have proposed to use generalised principal component analyses with convenient choices of the metric on the units space. We first recall their proposal for the graphical detection of outliers. Then we show how to improve it by using recent theoretical results by Hakam (2002) : these results allow us to get a test to assess the statistical signification of the obtained projections, i.e. in practice to make a decision about the dimensionality of the relevant display. A simulated and a real examples are provided to illustrate the method and investigate its efficiency.

Keywords : *Graphical displays, Outliers, Principal component analysis, Projection pursuit.*

1. Introduction

Nous considérons n individus caractérisés par p variables réelles. A chaque individu correspond donc un vecteur X_i de $\mathbb{R}^p$ $(i = 1, \ldots, n)$, l'empilement de ces vecteurs conduisant à une matrice X de dimension $n \times p$. La méthode d'analyse (et de visualisation) de X la plus classique est l'Analyse en Composantes Principales (ACP) avec ses diverses variantes (ACP réduite, Analyse des Correspondances, etc.) qui la modifient le plus souvent par une transformation simple des données ou en spécifiant une métrique sur $\mathbb{R}^p$. Pour une discussion de tels choix on peut, par exemple, se référer à Besse, Caussinus, Ferré, Fine (1986). Ces diverses méthodes d'analyse restent cependant "du second ordre" en ce sens qu'elles sont fondées sur une inertie ou, si l'on préfére, sur des moments d'ordre 2. Elles ont été critiquées pour cela, ou plutôt leur efficacité a été relativisée (Sibson, 1984) car, de ce fait, certains aspects importants de la structure du nuage des X_i peuvent complètement leur échapper, ne pas apparaître dans les sous-espaces principaux retenus, même si leur dimension est élevée. Afin de mettre en évidence de telles structures cachées, sont nées des techniques dites ici de "projections révélatrices", heureuse traduction (due à Yves Escoufier) de l'anglais "Projection Pursuit" qui vient de la référence à une suite de projections à la recherche des aspects cachés du nuage des individus (Friedman et Tukey, 1974). Pour une formalisation synthétique on pourra consulter Huber (1985) et, pour une présentation en français, le chapitre 8 de l'ouvrage collectif édité par Droesbecke, Fichet et Tassi (1992). Par ailleurs, alors que la plupart des travaux sur les projections révélatrices sont fondés sur des approches très différentes de celle de l'ACP, ce dernier texte met l'accent sur la possibilité d'obtenir des projections originales intéressantes par des méthodes proches de l'ACP, en ce sens qu'elles se ramènent à la simple diagonalisation d'une matrice : il s'agit en fait d'ACP "généralisées" utilisant sur $\mathbb{R}^p$ des métriques convenables dépendant des données (de façon non quadratique si bien que les méthodes associées ne sont plus seulement du second ordre). Ces techniques d'analyse ont ultérieurement été développées par Caussinus et Ruiz-Gazen (1993, 1995) qui montrent quelques propriétés des méthodes d'ACP généralisée qu'ils ont proposées. On trouvera en particulier dans ces articles un cadre théorique permettant de fournir des réponses à des questions telles que : pour quel type de structure des données ces méthodes sont-elles susceptibles de produire des projections intéressantes? Les problèmes de dimension y sont aussi abordés : quelle est la dimension utile de l'espace de projection (suffisamment grande pour capter la structure cherchée, suffisamment petite pour ne pas exhiber d'éventuels artefacts)? La solution à cette dernière question, telle que proposée dans ces articles, est largement heuristique. Mais des travaux récents (Hakam, 2002) permettent de la préciser, de sorte que c'est une démarche plus achevée que nous pouvons maintenant mettre en place et que nous nous proposons de présenter ici en nous attachant aux aspects méthodologiques (nous renvoyons ailleurs pour des démonstrations d'ordre mathématique) et en les illustrant par des exemples.

Dans le présent article, nous nous concentrons sur la détection et la visualisation des individus atypiques. Si le problème des valeurs atypiques a suscité une abondante littérature, c'est bien davantage pour les données unidimensionnelles que pour les données multidimensionnelles. Le concept d'individu atypique est alors nettement plus complexe et les techniques graphiques ont une importance majeure pour leur mise en évidence. Ces techniques restent le plus souvent heuristiques et non contrôlées par

des outils de validation (voir, par exemple, Barnett et Lewis, 1994, section 7.4). Au contraire, la démarche proposée dans cet article relève à la fois de l'analyse descriptive des données et d'une validation de nature confirmative en nous permettant :

(i) d'une part de détecter les individus suspects de s'éloigner du nuage "majoritaire" et de les "positionner" par rapport à celui-ci,

(ii) d'autre part de décider lesquels de ces individus, ou du moins quelles projections les mettant en évidence, sont significatif(ve)s, par rapport à ce qui résulte plus vraisemblablement d'un seul bruit fortuit.

Sans doute la possibilité de réunir naturellement ces deux aspects de l'analyse statistique constitue-t-elle l'intérêt majeur de la méthode globale que nous présentons ci-dessous, le point (i) dans le paragraphe 2, le point (ii) dans le paragraphe 3, tandis que des exemples viennent illustrer le tout dans le paragraphe 4.

2. ACP généralisée pour la recherche d'individus atypiques : modèle et principe

A chaque individu i ($i = 1, \ldots, n$) est associé un vecteur X_i de $\mathbb{R}^p$. On considère les deux matrices $p \times p$ suivantes :

$$V_n = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)(X_i - \bar{X}_n)'$$

$$S_n(\beta) = \frac{\sum_{i=1}^{n} \exp\left(-\frac{\beta}{2} \left\|X_i - \bar{X}_n\right\|^2_{V_n^{-1}}\right) (X_i - \bar{X}_n)(X_i - \bar{X}_n)'}{\sum_{i=1}^{n} \exp\left(-\frac{\beta}{2} \left\|X_i - \bar{X}_n\right\|^2_{V_n^{-1}}\right)} \qquad (1)$$

avec $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$ et $\|X\|^2_M = X' M X$.

V_n est la matrice des variances et covariances empiriques usuelles, tandis que $S_n(\beta)$ peut s'interpréter comme une matrice de variances et covariances robustes (on peut consulter Ruiz-Gazen, 1996, pour des précisions sur cet aspect) dans laquelle β est un réel donné "petit" (voir ci-dessous).

Selon une idée initiée par Yenyukov (1988) (sous une forme heuristique et un peu différente) et développée par Caussinus et Ruiz-Gazen (1993, 1995), on réalise une ACP de V_n avec, pour métrique sur l'espace $\mathbb{R}^p$ des individus, la matrice inverse de $S_n(\beta)$. On diagonalise donc $V_n S_n^{-1}(\beta)$ et les individus sont projetés $S_n^{-1}(\beta)$-orthogonalement sur les sous-espaces principaux. Il est important de rappeler que la projection ainsi obtenue est invariante par transformation affine des X_i, ce qui fait bien ressortir qu'elle concerne seulement la structure du nuage des individus au delà des divers aspects de centrage et d'échelle (même pour des changements d'échelle arbitrairement choisis pour *chacune* des variables, alors que l'ACP usuelle n'est invariante que par transformations orthogonales, c'est-à-dire préservant les distances).

Considérons maintenant le modèle suivant de "valeurs atypiques" :

X_i est un vecteur aléatoire dont la loi de probabilité est un mélange de $(q+1)$ lois normales $\mathcal{N}(m_j, W)$ en proportions respectives p_j $(j = 0, 1, \ldots, q)$, où l'indice 0 caractérise la loi majoritaire et les indices 1 à q caractérisent q possibilités de contamination de la moyenne. En vertu de la propriété d'invariance rappelée ci-dessus, on peut supposer $m_0 = 0$ et $W = I$.

Pour faire ressortir au mieux les valeurs atypiques sur les sous-espaces de projection, il faut que la méthode conduise à projeter les individus sur le sous-espace engendré par les m_j $(j = 1, \ldots, q)$. On peut montrer (voir Caussinus et Ruiz-Gazen, 1993, 1995) que c'est le cas, pour n assez grand et des p_j $(j = 1, \ldots, q)$ suffisamment petits (rappelons que l'objectif est de mettre en évidence des individus atypiques, donc des configurations rares), en prenant pour β une valeur petite, par exemple de l'ordre de 0.05 ou 0.01. Pour des précisions sur les aspects théoriques, on pourra consulter les deux articles mentionnés ci-dessus. Par ailleurs, le nombre q de contaminations est, en pratique, inconnu. Il est égal à la dimension du sous-espace engendré par les m_j $(j = 1, \ldots, q)$, c'est-à-dire q (sauf dans le cas exceptionnel où les m_j en question ne seraient pas linéairement indépendants). Ainsi, le choix de la dimension du sous-espace utile pour visualiser les valeurs atypiques est-il lié à l'évaluation de l'entier inconnu q. Une méthode heuristique est proposée par Caussinus et Ruiz-Gazen dans les articles déjà cités. On remarque que, sous des conditions générales précisées par ces auteurs, les q plus grandes valeurs propres de l'ACP proposée, c'est-à-dire celles de la matrice $V_n S_n^{-1}(\beta)$, convergent, pour n grand, vers un nombre strictement supérieur à $1 + \beta$, tandis que les valeurs propres suivantes convergent vers $1 + \beta$. Cela conduit à arrêter l'ACP lorsque les valeurs propres de $V_n S_n^{-1}(\beta)$ sont proches de $1 + \beta$. Une seule question concrète se pose alors : que doit-on entendre par "proche de $1 + \beta$"? Cela dépend évidemment de la variabilité d'échantillonnage des valeurs propres considérées. Or, des travaux récents (Hakam, 2002) permettent de fournir une réponse de nature objective que nous exposons ci-dessous.

3. Choix de la dimension : valeurs critiques

Comme nous venons de le voir, le problème de la dimension est lié à l'évaluation de la loi de probabilité des valeurs propres de la matrice :

$$L = V_n S_n^{-1}(\beta).$$

Celles-ci sont liées de façon simple au valeurs propres de la matrice

$$M = \frac{\sqrt{n}}{\beta}\left(I - (1+\beta)V_n^{-1/2} S_n(\beta) V_n^{-1/2}\right).$$

Cette dernière matrice est symétrique. En notant M^* le vecteur colonne à $p(p+1)/2$ lignes constitué (dans l'ordre) des éléments diagonaux de M et des éléments du triangle supérieur, Hakam (2002) a obtenu le résultat suivant.

En l'absence de contamination, c'est-à-dire sous l'hypothèse nulle $q = 0$, la loi de M^* converge, quand n tend vers l'infini, vers une loi normale de moyenne nulle,

dont la matrice des variances et covariances a la forme

$$\left(\begin{array}{cc|cc} b & d & & \\ & \ddots & & 0 \\ d & b & & \\ \hline & & c & 0 \\ & 0 & & \ddots \\ & & 0 & c \end{array}\right)$$

où le bloc sud-est est diagonal de dimension $p(p-1)/2$ par $p(p-1)/2$ avec sur la diagonale

$$c = \frac{(\beta+1)^{p+2}}{\beta^2(2\beta+1)^{p/2+2}} - \frac{1}{\beta^2(\beta+1)^2},$$

et où le bloc nord-ouest est de dimension p par p avec tous les éléments extra-diagonaux égaux à

$$d = \frac{(\beta+1)^p}{(2\beta+1)^{p/2+2}},$$

et les éléments diagonaux égaux à

$$b = 2c + d.$$

Le terme c correspond à la variance asymptotique des termes extra-diagonaux de M, les covariances correspondantes étant nulles, tandis que b correspond à la variance asymptotique et d à la covariance asymptotique des termes diagonaux de M. Les covariances asymptotiques entre termes diagonaux et termes extra-diagonaux de M sont toutes nulles.

Des simulations ont montré que la convergence est relativement rapide et que l'approximation par la loi indiquée est bonne dès que n est voisin de 100.

A partir de là, la loi asymptotique des valeurs propres n'est pas facile à déterminer de façon analytique, mais il est simple et rapide de le faire par simulation puisqu'il suffit de générer des matrices dont la loi est la loi limite donnée ci-dessus, matrices très facilement obtenues par transformations élémentaires de lois $\mathcal{N}(0,1)$.

Cela permet de proposer une technique de validation contrôlée du choix de la dimension, par exemple un test d'hypothèses multiples défini ainsi :

- choisir un niveau de signification α. La table 1 correspond à $\alpha = 5\%$ et la table 2 à $\alpha = 1\%$.
- transformer les valeurs propres de L en les valeurs propres correspondantes de M. On a :

si λ est valeur propre de L, alors $\mu = \frac{\sqrt{n}}{\beta}\left[1 - \frac{\beta+1}{\lambda}\right]$ est valeur propre de M (les valeurs propres de L et M étant rangées dans le même ordre),

- comparer chaque valeur propre empirique ainsi obtenue à la valeur critique correspondante en commençant par la plus grande valeur propre et en validant les dimensions comme significatives tant que la valeur empirique des valeurs propres de M est supérieure à la valeur critique.

Remarquons que les valeurs critiques calculées dépendent du paramètre β. Dans Caussinus et Ruiz-Gazen (1993, 1995), les propriétés théoriques ont été développées pour β suffisamment petit et, dans les exemples analysés, β était choisi de l'ordre de 0.1 à 0.5. Nous suggérons d'utiliser plutôt des valeurs encore inférieures pour β, de l'ordre de 0.01 à 0.1. Les résultats obtenus sont particulièrement stables pour ces différentes valeurs de β. Les valeurs critiques des tables 1 et 2 ont été évaluées par simulation comme indiqué plus haut pour $\beta = 0.05$ et sont valides pour β compris entres 0.01 et 0.1 avec une précision de l'ordre de 0.1 (demi-longueur approximative des intervalles de confiance de sécurité 0.95). Nous utilisons la valeur $\beta = 0.05$ dans les exemples traités ci-dessous.

Comme toujours, dans une règle de choix multiple, il est clair que le niveau α n'est qu'indicatif au delà de la dimension 1. Cette question pourrait mériter plus ample discussion, mais il nous semble qu'en matière de contrôle d'une expression graphique, nous avons déjà ainsi un cadre suffisamment solide.

4. Exemples

Nous examinons deux exemples. Le premier est un exemple simulé précédemment analysé dans Caussinus et Ruiz-Gazen (1993). Il s'agit d'un échantillon de 100 observations à 6 dimensions dont les 10 premières sont issues de la loi $\mathcal{N}((5,0,0,0,0,0)', W)$ et les 90 suivantes sont issues de la loi normale $\mathcal{N}(0, W)$ avec W matrice diagonale de diagonale $(1,4,4,4,4,4)$. D'après la distribution de l'échantillon, $q = 1$. L'ACP généralisée avec la métrique $S_n^{-1}(0.05)$ conduit aux valeurs propres de L et M suivantes :

	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
L	1.12	1.07	1.05	1.05	1.02	1.01
	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6
M	12.15	3.21	0.44	−0.78	−5.93	−8.29

D'après la table 1, on trouve qu'au niveau $\alpha = 5\%$, seule la première dimension est à retenir pour visualiser les individus atypiques. La figure 1 montre qu'effectivement les 10 premières observations forment un nuage à peu près homogène sur la droite du graphique qui se distingue bien (sauf l'observation 4) du reste des données sur le premier axe. Sans critère de validation, il aurait été tentant d'affirmer que l'observation numéro 13 était aussi atypique, dans une autre direction

TABLE 1
Valeurs critiques des 10 plus petites valeurs propres
de M pour $p = 1, \ldots, 20$ au niveau $\alpha = 5\%$

Valeurs critiques										
p	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_7	μ_8	μ_9	μ_{10}
2	5.5	1.5								
3	7.1	3.2	−0.3							
4	8.4	4.7	1.6	−1.6						
5	9.9	6.0	3.0	0.2	−2.9					
6	11.5	7.3	4.4	1.7	−1.1	−4.1				
7	12.8	8.8	5.7	2.9	0.4	−2.3	−5.4			
8	14.4	10.1	7.0	4.2	1.6	−0.9	−3.5	−6.5		
9	15.9	11.3	8.1	5.4	2.9	0.5	−1.9	−4.5	−7.6	
10	17.3	12.7	9.5	6.7	4.1	1.7	−0.7	−3.2	−5.8	−9.0
11	18.8	14.0	10.7	7.8	5.3	2.9	0.6	−1.7	−4.2	−6.9
12	20.0	15.3	12.0	9.1	6.6	4.1	1.7	−0.5	−2.9	−5.4
13	21.8	16.7	13.2	10.2	7.6	5.2	2.9	0.7	−1.6	−4.0
14	23.1	18.0	14.5	11.5	8.9	6.4	4.0	1.8	−0.4	−2.7
15	24.6	19.3	15.7	12.6	9.9	7.4	5.1	2.8	0.8	−1.5
16	25.8	20.6	17.0	13.9	11.2	8.6	6.3	4.0	1.8	−0.4
17	27.4	22.0	18.2	15.0	12.2	9.7	7.4	5.1	2.9	0.7
18	28.7	23.3	19.5	16.4	13.5	10.9	8.5	6.2	4.0	1.8
19	30.3	24.7	20.8	17.5	14.7	12.0	9.6	7.3	5.1	2.9
20	31.5	25.9	22.0	18.7	15.9	13.2	10.8	8.4	6.2	4.0

que les précédentes; mais le fait que la seconde direction principale ne soit pas significative prévient de s'égarer vers cette interprétation malencontreuse du graphique; 13 est certes extrême, mais tout à fait compatible avec la variabilité du groupe principal.

Une technique classique de détection de valeurs atypiques en situation multivariée consiste à considérer comme telle toute observation X_i dont la distance de Mahalanobis au centre de la distribution $(X_i - \bar{X}_n)' V_n^{-1} (X_i - \bar{X}_n)$ excède une valeur critique donnée. Rousseeuw et Van Zomeren (1990) ont proposé d'utiliser une version robuste de la distance de Mahalanobis en remplaçant $\bar{X}_n$ et V_n par des estimateurs de position et de dispersion robustes. Dans Ruiz-Gazen (1996), il est proposé d'utiliser un estimateur simple et robuste de dispersion défini par : $U_n(\beta) = (S_n^{-1}(\beta) - \beta V_n^{-1})^{-1}$. Cet estimateur de dispersion nous semblant le plus cohérent avec la méthode de projections révélatrices utilisée ici, nous avons calculé les distances de Mahalanobis des 100 observations de l'exemple (distances à la moyenne au sens de la métrique

TABLE 2
Valeurs critiques des 10 plus petites valeurs propres
de M pour $p = 1, \ldots, 20$ au niveau $\alpha = 1\%$

Valeurs critiques										
p	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_7	μ_8	μ_9	μ_{10}
2	7.1	2.9								
3	8.7	4.4	1.1							
4	10.1	6.1	2.9	−0.3						
5	11.6	7.3	4.3	1.4	−1.5					
6	13.2	8.6	5.7	2.9	0.1	−2.7				
7	14.7	10.3	7.0	4.1	1.6	−1.0	−3.9			
8	16.2	11.5	8.4	5.5	2.9	0.4	−2.1	−5.0		
9	18.0	12.9	9.5	6.6	4.1	1.8	−0.8	−3.2	−6.1	
10	19.2	14.3	11.0	8.1	5.4	3.1	0.6	−1.9	−4.5	−7.4
11	21.1	15.6	12.1	9.2	6.5	4.2	1.8	−0.4	−2.9	−5.6
12	22.4	17.0	13.3	10.5	7.9	5.3	3.0	0.7	−1.7	−4.2
13	24.3	18.4	14.7	11.6	8.9	6.5	4.0	1.9	−0.5	−2.7
14	25.4	19.7	15.9	12.8	10.2	7.6	5.1	3.0	0.7	−1.5
15	27.2	21.2	17.5	14.1	11.4	8.9	6.5	4.2	2.1	−0.2
16	28.3	22.5	18.4	15.4	12.5	9.9	7.5	5.2	3.0	0.8
17	30.3	23.9	20.0	16.6	13.6	10.9	8.6	6.3	4.2	2.0
18	31.4	25.1	21.1	17.9	14.9	12.3	9.8	7.4	5.2	2.9
19	33.3	26.6	22.4	19.0	15.9	13.2	10.9	8.6	6.2	4.1
20	34.2	28.1	23.9	20.3	17.3	14.6	12.0	9.6	7.4	5.2

$U_n^{-1}(0.05)$). Bien entendu les valeurs décidées atypiques dépendent du niveau de signification retenu (sans même parler des problèmes de niveau posés par la réalisation de tests successifs). Ainsi, les individus numéros 9, 1, 6, 8, 10, 5, sont les seuls parmi les "vrais" atypiques à être détectés à un niveau de l'ordre de 5%, mais en même temps que l'on décrète atypiques les individus "normaux" 13 et 69. En allant jusqu'à un niveau de 10%, on retrouve l'individu 3 (mais aussi 26, 15 et 72!), au niveau 20% les individus 2 et 7 sont détectés (mais aussi 85, 79, 51, 33, 99!), etc. Quel que soit le niveau choisi, les distances de Mahalanobis calculées sur l'ensemble des variables ne permettent pas de bien distinguer les "vrais" individus atypiques des autres individus. Remarquons que ce constat reste vrai si l'on remplace la moyenne et l'estimateur $U_n(\beta)$ par des estimateurs à haut point de rupture (voir Rousseeuw et Van Zomeren, 1990).

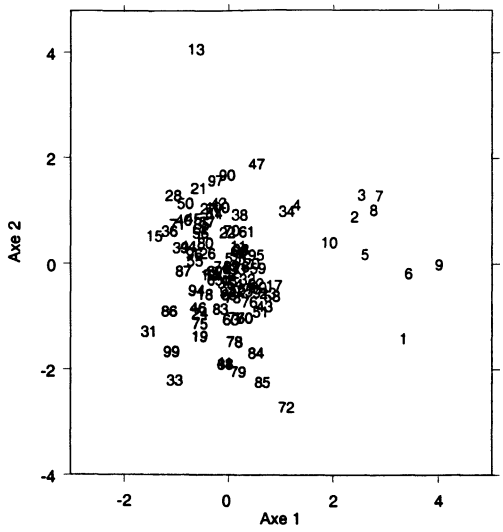

FIGURE 1
Premier plan principal de l'ACP avec la métrique $S_n^{-1}(0.05)$
pour l'exemple simulé

Une fois choisie la dimension de représentation, on peut aussi suggérer de calculer des distances de Mahalanobis dans l'espace de projection des individus. En fait, sur les représentations graphiques que nous donnons, la distance entre points correspond à la distance de Mahalanobis au sens de la métrique $S_n^{-1}(\beta)$. Si l'utilisateur fixe un niveau de signification α, on peut tracer un cercle de confiance associé dont le rayon correspond à la racine carrée du quantile d'ordre $1-\alpha$ d'une loi de χ^2 à 2 degrés de liberté multipliée par la racine carrée de $1+\beta$. Le coefficient $1+\beta$ est justifié car, pour des observations gaussiennes de matrice de covariances W, $(1+\beta)S_n(\beta)$ est un estimateur consistant de W. Cela n'a pas été fait sur la figure 1 pour éviter d'alourdir, d'autant que tracer ici un cercle peut être trompeur, incohérent avec le fait qu'une seule dimension est significative. Par contre, dans l'exemple qui suit, nous traçons, pour chaque plan principal considéré, deux cercles centrés au point moyen et dont les rayons correspondent respectivement à un niveau de 1% et de 5%.

Le deuxième exemple est extrait de Daudin, Duby et Trécourt (1988). Il s'agit de $p = 8$ mesures effectuées sur 86 échantillons de lait. Ces données contiennent un certain nombre d'observations atypiques. L'observation 70 en particulier est extrême pour la cinquième et la sixième variable qui valent respectivement 33.8 et 33.7 pour cette observation alors que les valeurs de la cinquième (respectivement la sixième) variable sont comprises entre 22.2 et 27.7 (respectivement 22.3 et 27.4) pour le reste des observations. Il s'agit très vraisemblablement d'une erreur de transcription et, dans Caussinus et Ruiz-Gazen (1995) comme ici, l'observation 70 a été corrigée en remplaçant 33.8 (respectivement 33.7) par 23.8 (respectivement 23.7).

L'ACP avec la métrique $S_n^{-1}(\beta)$ permet de repérer plusieurs observations atypiques. Atkinson (1994) a repris l'exemple en enlevant l'observation 70 et en

utilisant la méthode du "stalactite plot". Cette méthode, proposée par Atkinson et Mulira (1993) est un raffinement de la méthode des distances de Mahalanobis : il s'agit d'un résumé graphique qui permet de détecter des valeurs atypiques au sein d'un échantillon en calculant des distances de Mahalanobis à partir de sous-échantillons non contaminés de taille croissante. Bartkowiak et Szustalewicz (2000) ont aussi considéré cet exemple. Ils ont conservé l'observation 70 et ont utilisé la méthode du "grand tour" pour détecter les observations atypiques dans ce fichier de données. Cette méthode, proposée par Asimov (1985), consiste en une suite de rotations et de projections des données en deux dimensions et repose sur l'utilisation d'un environnement graphique dynamique. Bartkowiak et Szustalewicz (2000) proposent de représenter sur ces suites de graphiques des ellipsoïdes de confiance en suspectant comme atypiques les observations qui se trouvent "fréquemment" à l'extérieur des ellipsoïdes.

En réalisant une ACP généralisée avec la métrique $S_n^{-1}(0.05)$, après correction de l'observation 70, on obtient les valeurs propres de L et M suivantes :

	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8
L	1.40	1.26	1.13	1.11	1.07	1.05	1.01	0.99
	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_7	μ_8
M	46.68	30.87	13.02	10.80	3.79	0.11	-8.17	-11.41

D'après les valeurs critiques précisées dans les tables 1 et 2, 5 dimensions sont à retenir au niveau de 1% et 6 dimensions au niveau de 5%. Les figures 2, respectivement 3 et 4, représentent les 86 observations projetées respectivement sur les plans principaux (1,2), (3,4) et (5,6) (d'autres combinaisons des 6 premiers axes n'apportent pas d'information supplémentaire). Ces représentations graphiques permettent non seulement de repérer les individus atypiques mais aussi de les cartographier. Comme expliqué plus haut, figurent sur les graphiques, à titre indicatif, deux cercles correspondant pour le plus petit au niveau $\alpha = 5\%$ et pour le plus grand au niveau $\alpha = 1\%$. Nous repérons ainsi sur la figure 2 qu'au niveau 1%, les individus 1, 2 et 41 sont atypiques sur le premier axe tandis que l'individu 44 et, dans une moindre mesure, l'individu 3 se distinguent de la majorité des données sur le deuxième axe. Sur cette figure, au niveau 5%, les individus 18, 27 et 75 peuvent aussi être suspectés atypiques. Sur la figure 3, au niveau 1%, les individus 12, 13, 14, 15, 47 et 20 se distinguent de la majorité des données (ainsi que 11 et 75 au niveau 5%). Enfin, sur la figure 4, au niveau 1%, nous repérons comme atypiques les individus 77 et 17 et, en plus, au niveau 5%, 16, 42, 75 et 85. Si nous comparons nos résultats avec ceux obtenus d'une part par Atkinson (1994) et d'autre part par Bartkowiak A. et Szustalewicz A. (2000), nous pouvons remarquer que nous avons détecté en commun avec ces auteurs les observations 1, 2, 12, 13, 14, 15, 16, 17, 41, 44, 77. Nous trouvons aussi, comme Atkinson (et contrairement à Bartkowiak et Szustalewicz), que les observations 3, 47 et 75 sont atypiques et comme Bartkowiak et Szustalewicz et (contrairement à Atkinson) que les observations 11 et 42 sont atypiques. La seule observation considérée comme atypique à la fois par Atkinson et Bartkowiak-Szustalewicz et que nous ne retrouvons pas au niveau des 3 graphiques "significatifs" présentés est l'observation 74. Considérant maintenant la figure 5 qui

correspond au plan principal (7,8), on peut noter qu'elle présente un nuage de points sans réelle structure conformément au fait qu'elle est associée à des dimensions non significatives; l'observation 74 (comme d'ailleurs 28, 35 et 49) y apparaît certes extrême mais sans que cela soit suffisant, d'après nous, pour la juger atypique (il est tout à fait naturel que 4 observations sur 86 se trouvent à l'extérieur du cercle 5%).

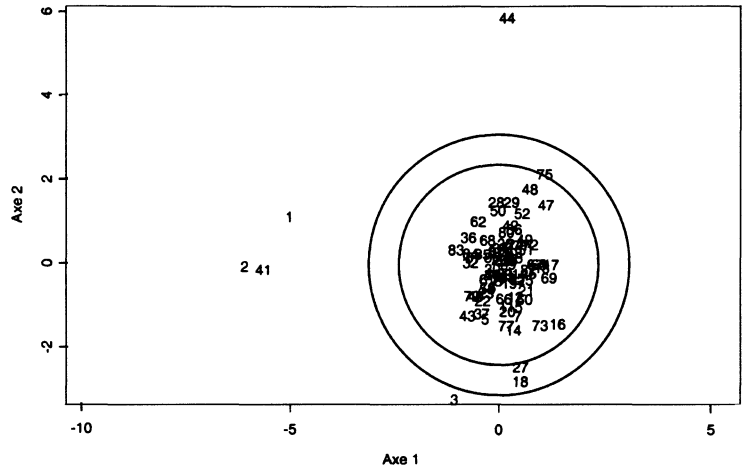

FIGURE 2
Plan principal (1,2) de l'ACP avec la métrique $S_n^{-1}(0.05)$
pour l'exemple du lait

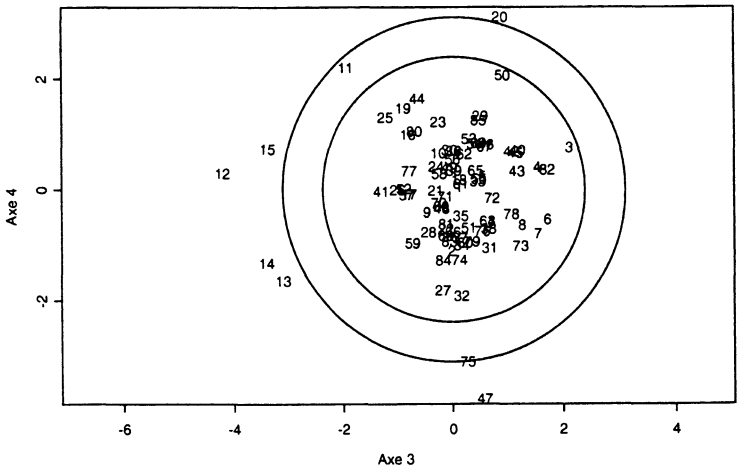

FIGURE 3
Plan principal (3,4) de l'ACP avec la métrique $S_n^{-1}(0.05)$
pour l'exemple du lait

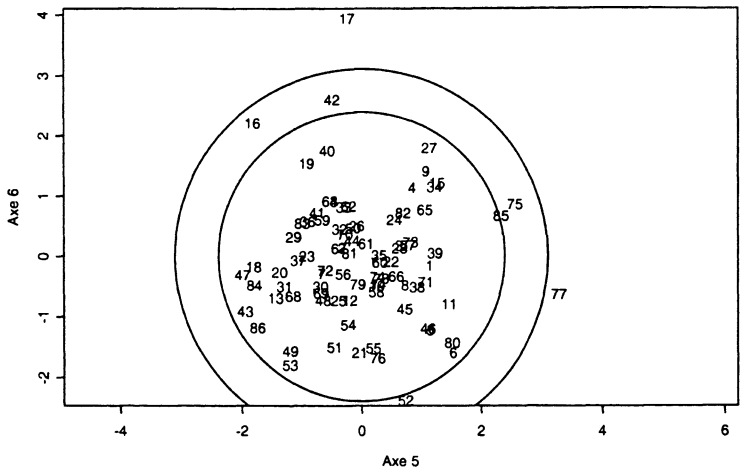

FIGURE 4
Plan principal (5,6) de l'ACP avec la métrique $S_n^{-1}(0.05)$ pour l'exemple du lait

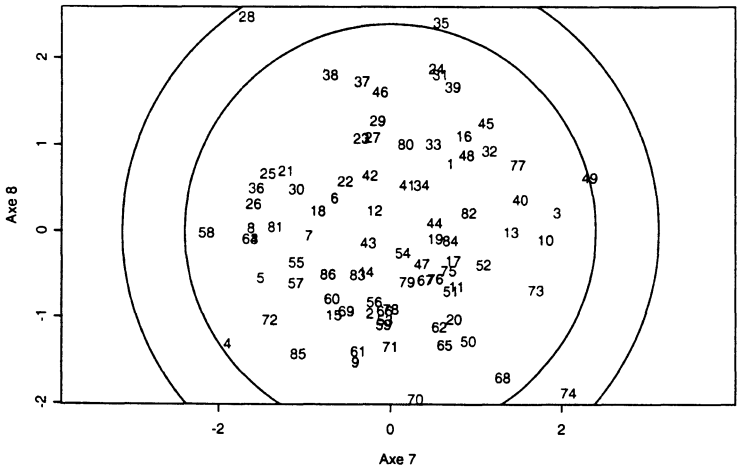

FIGURE 5
Plan principal (7,8) de l'ACP avec la métrique $S_n^{-1}(0.05)$ pour l'exemple du lait

5. Conclusion

Si le problème de détection des valeurs atypiques est ancien, il a reçu récemment une attention particulière dans le cadre du "Data Mining" dont une présentation

synthétique et critique pourra être trouvée, par exemple, dans l'article de Besse, Le Gall, Raimbault et Sarpy (2001). Cependant, comme le rappellent ces auteurs, si le chercheur d'or ne trouve pas de pépite là où il n'y en pas, en fouillant suffisamment les données, le statisticien peut très bien "découvrir" des pépites illusoires. C'est la raison pour laquelle nous croyons important de contrôler efficacement toute méthode d'exploration, de "fouille", des données qui, comme une technique de projections révélatrices, cherche à mettre en évidence quelque structure spécifique cachée éventuelle, ici l'existence d'individus atypiques. Le problème de la détection surabondante de valeurs atypiques est d'ailleurs souvent évoqué dans la littérature (voir par exemple Atkinson, 1994 et Bartkowiak et Szustalewicz, 2000).

Sur les exemples présentés, on voit les avantages et les limites de la méthode de contrôle proposée. Pour les avantages, on peut souligner à nouveau le fait de ne pas affirmer (à tort) que l'observation 13 du premier exemple est atypique et de résister ainsi à une illusion à laquelle il serait difficile d'échapper autrement. Les inconvénients de la méthode sont évidemment duaux de ses avantages : tout contrôle peut être excessif (c'est l'éternelle question des erreurs de première et de seconde espèce). En fait, dans cet article, nous n'avons pas considéré les problèmes de puissance. Disons cependant brièvement que celle-ci est une fonction croissante, d'une part de la distance des valeurs atypiques au groupe majoritaire (distance mesurée au travers de $m_j - m_0$), d'autre part de la probabilité p_j ; de façon qui peut paraître paradoxale, si p_j doit être petite pour que la méthode de projection fonctionne (cf. Caussinus et Ruiz-Gazen, 1993), la probabilité de détection sera plus faible si p_j est trop petite : pratiquement, il est plus facile de détecter un groupe d'individus atypiques semblables qu'un seul individu, ce qui est finalement naturel car un groupe "pèse" davantage dans la recherche de la direction qui le caractérise. Dans notre second exemple, c'est peut-être la raison pour laquelle l'individu 74 n'est pas jugé atypique (mais une autre raison possible est simplement qu'il ne l'est pas; dans cet exemple réel, il est évidemment impossible de savoir de façon certaine).

Déterminer la dimension significative q des données, c'est dégager les aspects importants par rapport au simple bruit. Ayant donc estimé q par la méthode proposée et projeté les individus dans le sous-espace principal correspondant à q dimensions, on peut mettre en œuvre, sur des données de plus petite dimension sans pour autant perdre une information substantielle, toute (autre) méthode de recherche de structure, par exemple ici, la méthode des "stalactite plots" ou la méthode du "grand tour" précédemment citées. Cette dernière pourrait avoir l'avantage de ne pas limiter les projections à des sous-espaces définis pas des axes factoriels, tandis que la précédente peut fournir des représentations graphiques complémentaires utiles.

6. Références

ASIMOV D. (1985), "The Grand Tour : a tool for viewing multidimensional data", SIAM J. Sci. Stat. Comput., vol.6, 1, 128-143.

ATKINSON A.C. (1994), "Fast very robust methods for the detection of multiple outliers", J. Amer. Statist. Assoc., 89, 1329-1339.

ATKINSON, A.C., MULIRA, H.-M. (1993), "The Stalactite Plot for the Detection of Multivariate Outliers", Statistics and Computing, 3, 27-35.

BARTKOWIAK A., SZUSTALEWICZ A. (2000), "Outliers - finding and classifying which genuine and which spurious", Comput. Statist., 15, 1, 3-12.

BARNETT V., LEWIS T. (1994), "Outliers in Statistical Data", 3rd edition, Wiley, New York.

BESSE Ph., CAUSSINUS H., FERRÉ L., FINE J. (1986), "Some guidelines for principal component analysis", COMPSTAT 86, Physica-Verlag, Heidelberg, 23-30.

BESSE Ph., LE GALL C., RAIMBAULT N., SARPY S. (2001), "Data Mining et Statistique" (avec discussion), Journal de la Société Française de Statistique, 142, 1, 5-95.

CAUSSINUS H., RUIZ-GAZEN A. (1993), "Projection Pursuit and Generalized Principal Component Analyses", In New Directions in Statitical Data Analysis and Robustness, Eds. Morgenthaler S. et al., Birkhäuse Verlag, Basel Boston Berlin, 35-46.

CAUSSINUS H., RUIZ-GAZEN A. (1995), "Metrics for finding typical structures by means of principal component analysis", In Data Science and its Applications, Eds. Escoufier Y. et al., Academic Press, 177-192.

DAUDIN J.J., DUBY C., TRÉCOURT P. (1988), "Stability of Principal Component Analysis by bootstrap method", Statistics, 19, 241-258.

DROESBECKE J.-J., FICHET B. et TASSI Ph. (1992), Modèles pour l'analyse des données multidimensionnelles, Economica, Paris.

FRIEDMAN J.H., TUKEY J.W. (1974), "A Projection Pursuit Algorithm for Exploratory Data Analysis", IEEE Trans. Comput., C-23, 881-889.

HAKAM S. (2002), Tests de signification pour quelques méthodes de projections révélatrices et applications, Thèse, Université Mohammed V-Agdal, Rabat.

HUBER P. J. (1985), "Projection Pursuit" (with discussion), Ann. Statist., 13, 435-525.

RUIZ-GAZEN A. (1996), "A very simple robust estimator for a dispersion matrix", Comput. Statist. and Data Anal., 21, 149-162.

ROUSSEEUW, P.J., VAN ZOMEREN, B.C. (1990), "Unmasking Multivariate Outliers and Leverage Points", Journal of the American Statistical Association, 85, 633-639.

SIBSON R. (1984), "Present position and potential developments : some personal views - multivariate analysis", J. R. Statist. Soc., A, 198-207.

YENYUKOV I.S. (1988), "Detecting structures by means of Projection Pursuit", COMPSTAT 88, Physica-Verlag, Heidelberg, 47-58.