REVUE DE STATISTIQUE APPLIQUÉE

R. CLEROUX
J.-M. HELBLING
N. RANGER

Détection d'ensembles de données aberrantes en analyse des données multivariées

Revue de statistique appliquée, tome 38, n° 1 (1990), p. 5-21
<http://www.numdam.org/item?id=RSA_1990__38_1_5_0>

© Société française de statistique, 1990, 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, 1990, XXXVIII (1), 5-21

DÉTECTION D'ENSEMBLES DE DONNÉES ABERRANTES EN ANALYSE DES DONNÉES MULTIVARIÉES

R. CLEROUX (1), J.-M. HELBLING (2) & N. RANGER (1)

(1) Département d'informatique et de recherche opérationnelle, Université de Montréal (Canada)
(2) Département de mathématiques, Ecole Polytechnique Fédérale de Lausanne (Suisse).

Recherche subventionnée en partie par le Conseil de recherche en sciences naturelles et en génie du Canada et par le Fonds National suisse pour la recherche scientifique.

RÉSUMÉ

Dans cet article on utilise la fonction d'influence d'un point sur le coefficient de corrélation vectorielle ρV pour déceler une donnée douteuse. La notion d'ensemble de points douteux est introduite et utilisée avec le concept de fonction d'influence. Pour procéder à la formation des groupes et pour éviter une énumération systématique, une méthode de classification est employée. Sur un exemple, la procédure proposée est appliquée et les résultats sont comparés à ceux obtenus par d'autres méthodes de détection de données aberrantes multivariées. Les concepts développés dans la première partie sont aussi mis en évidence dans le cadre de la régression linéaire multivariée par l'intermédiaire de la fonction d'influence du coefficient $\rho V reg$. Un exemple illustre la procédure dans ce cadre plus particulier.

Mots clés : *Données multivariées aberrantes, détection d'ensembles de points aberrants, fonction d'influence, corrélation vectorielle, classification.*

1. Introduction

En analyse des données comme d'ailleurs en statistique inférentielle multivariée, il est essentiel de travailler sur des données expérimentales fiables. Pour évaluer la qualité de son échantillon le statisticien doit disposer de techniques permettant de déceler des données douteuses ou franchement erronées. Ce genre de données se rencontre dans la littérature sous divers noms : données aberrantes, observations influentes, aberrations, valeurs surprises, valeurs extrêmes, etc...

Si en statistique univariée il suffit d'examiner les points dans les queues de la distribution échantillonnale pour tester la présence d'observations aberrantes, en situation multivariée cette détection requiert beaucoup plus de précau-

tion. En effet, qu'est-ce qu'une donnée aberrante dans un échantillon de $\mathbb{R}^m$ ($m > 1$)? Une définition rigoureuse est impossible à donner dans l'absolu car ce qui est aberrant pour un modèle ne l'est pas forcément pour un autre et de nombreux types différents d'aberrances peuvent survenir en statistique multivariée comme le souligne Gnanadesikan (1977, p. 271). Toutefois on peut dire que les données douteuses se trouveront à la périphérie du nuage de points formé par l'échantillon. On suit en cela la définition très générale de Grubbs (1969) "An outlying observation, or outlier, is one that appears to deviate markedly from the other members of the sample in which it occurs".

Pourquoi s'intéresser aux données douteuses? Les données aberrantes ont parfois un intérêt en elles-mêmes (en pollution par exemple), par ailleurs elles peuvent être la source de contamination dans des analyses futures, telles que dans l'ajustement d'un modèle, l'estimation de paramètres ou dans les tests d'hypothèses. Il est donc nécessaire de s'en préoccuper au début de toute étude.

La détection d'observations aberrantes multivariées est un problème complexe. L'utilisation de techniques univariées appliquées aux projections sur chaque axe ne conduit pas nécessairement à un bon résultat comme le laisse apparaître la figure 1.1. :

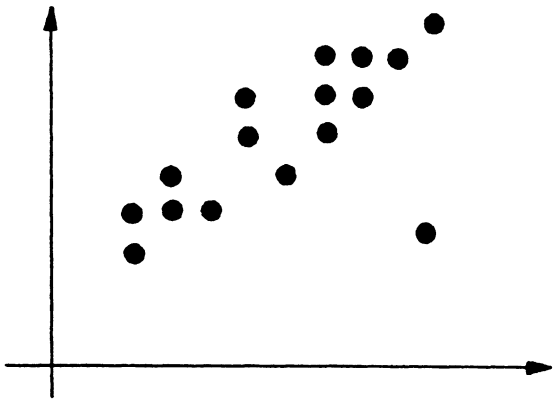


FIGURE 1.1.
Données artificielles dans le plan

Depuis vingt ans environ plusieurs méthodes de détection de données aberrantes ont été proposées. Beaucoup d'entre elles traitent le cas univarié. On peut lire à ce sujet les livres ou les articles de Barnett et Lewis (1984), Hawkins (1980), Jain (1981), Kale (1976, 1979), Prescott (1980) et Rosner (1983). Beckman et Cook (1983) présentent une excellente synthèse de ce problème. D'autres techniques sont liées spécifiquement au modèle linéaire comme celles présentées par Andrews et Pregibon (1978), Bradu et Hawkins (1982), Cook (1977) et (1979), Cook et Weisberg (1980), Draper, Guttman et John (1984), Gentleman (1980), Hawkins, Bradu et Kass (1984), Kotze et Hawkins (1984), Rousseeuw et Leroy (1987) et Schall et Dunne (1987).

En statistique multivariée, les méthodes proposées se basent généralement sur un modèle développé pour des populations normales. Des tests sont alors mis au

point pour la détection d'une aberration comme le rapportent Barnett (1979, 1983), Garel (1978), Guttman (1973), Hawkins (1974, 1980 Chap. 8) Siotani (1959), Schwager et Margolin (1982) et Barnett et Lewis (1984, Chap. 9). Rohlf (1975) et Wilks (1963) proposent des tests pouvant se généraliser à plusieurs observations aberrantes mais toujours dans le cas de distributions normales. Ces tests sont basés sur des statistiques dont les distributions peuvent être approchées par des lois connues; leurs puissances sont inconnues. Basée sur la méthode de Wilks, Bacon-Shone et Fung (1987) proposent une représentation graphique qui permet la détection de données aberrantes.

Quelques auteurs, Sinha (1984) et Das et Sinha (1986) proposent des méthodes pouvant se généraliser à des distributions elliptiquement symétriques. Notons que dans le cadre de l'utilisation de techniques robustes, Hampel (1985) et Rousseeuw et van Zomeren (1987) proposent des techniques de détection et effectuent des comparaisons avec d'autres méthodes connues dans la littérature.

Finalement certaines techniques de détection de données aberrantes multivariées sont des procédures heuristiques qui n'utilisent pas d'hypothèses distributionnelles. Il s'agit de méthodes très utiles en analyse exploratoire des données. Les articles principaux s'y rapportant sont ceux d'Andrews (1972), Campbell (1978), Chernoff (1973), Chernoff et Rizvi (1975), Devlin, Gnanadesikan et Kettenring (1975), Gnanadesikan et Kettenring (1972), Gnanadesikan (1977, Chap. 6), Teillard et Volle (1976).

Comme on l'a déjà dit au début de cette introduction, il est très difficile de définir ce qu'est une aberrance multivariée. Une donnée peut être identifiée comme une aberrance par une méthode et ne pas l'être par une autre. Dès lors nous pensons qu'en toutes circonstances le statisticien qui trouve une ou des données douteuses doit discuter avec l'expérimentateur et qu'ensemble ils répondent à la question : est-on en présence de réelles aberrances et si oui qu'en fait-on? A ce sujet McCulloch et Meeter (1983) font une intéressante discussion.

Dans cet article, nous généralisons les résultats obtenus par Cléroux, Helbling et Ranger (1986, 1989) sur la fonction d'influence de deux coefficients de corrélation vectorielle à la détection de groupes de données aberrantes. Dans le paragraphe 2, on définit l'influence d'un groupe de points sur le coefficient ϱV et on propose une méthode heuristique de détection d'ensembles de données douteuses. De plus, pour éviter une énumération systématique des groupements possibles, on propose l'utilisation d'une méthode de classification. Dans le paragraphe 3, on fait le même développement pour le coefficient $\varrho Vreg$ utilisé en régression linéaire multivariée. Dans chacun des deux paragraphes des exemples sont traités afin d'illustrer les techniques proposées.

2. Le coefficient ϱV : Fonction d'influence et données aberrantes

2.1. Le coefficient ϱV

Soit $X = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix}$ un vecteur aléatoire tel que $X^{(1)}$ est un vecteur $p \times 1$ et $X^{(2)}$ est $q \times 1$. Sa moyenne μ et sa matrice de variance-covariance $\sum$ sont :

$$\mu = E(X) = \begin{pmatrix} \mu^{(1)} \\ \mu^{(2)} \end{pmatrix} \text{ et } \sum = E[(X-\mu)(X-\mu)'] = \begin{pmatrix} \sum_{11} & \sum_{12} \\ \sum_{21} & \sum_{22} \end{pmatrix}$$

Escoufier (1973) a introduit le coefficient de corrélation vectorielle ϱV dont la définition est :

$$\varrho V = \varrho V\left(X^{(1)}, X^{(2)}\right) = \frac{tr\left(\sum_{12}\sum_{21}\right)}{\sqrt{tr\left(\sum_{11}^2\right) tr\left(\sum_{22}^2\right)}}$$

Les propriétés de ce coefficient sont les suivantes :

i) Si $p = q = 1$ alors $\varrho V = \varrho^2$ le carré du coefficient de corrélation simple.

ii) $O \leq \varrho V \leq 1$ et

1) $\varrho V = 0$ si et seulement si $\sum_{12} = 0$

2) $\varrho V = 1$ *si* $X^{(2)} = AX^{(1)} + b$ où A est une matrice $q \times p$ telle que $A'A = kI$ (k scalaire positif) et b est un vecteur $q \times 1$ (A *et* b constants).

iii) Si A est une matrice $p \times p$ telle que $A'A = kI$(**k scalaire positif**) alors $\varrho V(AX^{(1)}, X^{(2)}) = \varrho V(X^{(1)}, X^{(2)})$.

L'utilisation de ce coefficient en statistique multivariée est multiple comme l'ont montré Robert et Escoufier (1976).

En présence d'un échantillon $X_1, X_2, ..., X_n$ on définit le coefficient de corrélation vectorielle échantillonnale en remplaçant dans l'expression de ϱV les paramètres par les estimateurs habituels pour obtenir

$$RV = RV\left(X^{(1)}, X^{(2)}\right) = \frac{tr\,(S_{12}S_{21})}{\sqrt{tr\,(S_{11}^2)\,tr\,(S_{22}^2)}}$$

où :

$$S_{ij} = \frac{1}{n-1}\sum_{\alpha=1}^{n}\left(X_\alpha^{(i)} - \bar{X}^{(i)}\right)\left(X_\alpha^{(j)} - \bar{X}^{(j)}\right)' \; i, j = 1, 2$$

$\bar{X}^{(i)}$ et $\bar{X}^{(j)}$ désignant respectivement les vecteurs moyennes empiriques des $X_\alpha^{(i)}$ **et** $X_\alpha^{(j)} (1 \leq \alpha \leq n)$.

En définissant les matrices :

$$Y_1 = \left(X_1^{(1)} - \bar{X}^{(1)}, X_2^{(1)} - \bar{X}^{(1)}, ..., X_n^{(1)} - \bar{X}^{(1)}\right) : p \times n$$
$$Y_2 = \left(X_1^{(2)} - \bar{X}^{(2)}, X_2^{(2)} - \bar{X}^{(2)}, ..., X_n^{(2)} - \bar{X}^{(2)}\right) : q \times n$$

et en utilisant la norme $\| E \| = \sqrt{trE'E}$, Robert et Escoufier (1976) ont montré que :

$$\sqrt{2}\sqrt{1 - RV\left(X^{(1)}, X^{(2)}\right)} = \text{dist } (Y_1, Y_2)$$

$$\text{où dist } (Y_1, Y_2) = \left\| \frac{Y_1'Y_1}{\sqrt{tr\,(Y_1'Y_1)^2}} - \frac{Y_2'Y_2}{\sqrt{tr\,(Y_2'Y_2)^2}} \right\|$$

Ainsi, la proximité entre les deux matrices de données Y_1 et Y_2 indique que la position relative des n points dans $\mathbb{R}^p$ et dans $\mathbb{R}^q$ est semblable.

2.2. Influence d'un point

La fonction d'influence d'un point X sur un paramètre θ est devenue un outil classique de la statistique depuis son introduction par Hampel (1974). Plusieurs auteurs, dont Devlin et al (1975) et Chatterjee et Hadi (1986), en font usage dans la détection de données aberrantes.

Considérons $\theta = T(F)$ comme fonctionnelle de la fonction de répartition F de la variable aléatoire X et soit $\tilde{F} = (1 - \epsilon)F + \epsilon\delta_x$ une perturbation de F par δ_x. δ_x est une fonction de répartition qui attribue la probabilité 1 au point x. La fonction d'influence $I(x;\theta)$ est alors définie par

$$I(x;\theta) = \lim_{\epsilon \to 0} \left(\frac{\tilde{\theta} - \theta}{\epsilon} \right) \text{ où } \tilde{\theta} = T(\tilde{F}) \tag{2.1}$$

Cléroux, Helbling et Ranger (1986) ont obtenu la fonction d'influence $I(X;\rho V)$ d'un point X sur le coefficient de corrélation vectorielle ρV :

$$I(X;\rho V) = \rho V \left[\frac{2Z^{(1)'}\Sigma_{12}Z^{(2)}}{tr\,(\Sigma_{12}\Sigma_{21})} - \frac{Z^{(1)'}\Sigma_{11}Z^{(1)}}{tr\,(\Sigma_{11}^2)} - \frac{Z^{(2)'}\Sigma_{22}Z^{(2)}}{tr\,(\Sigma_{22}^2)} \right] \tag{2.2}$$

où $Z^{(i)} = X^{(i)} - \mu^{(i)}, i = 1, 2$. Il est aisé de voir que cette fonction peut s'exprimer comme une forme quadratique

$$I(X;\varrho V) = Z'A_1Z \tag{2.3}$$

avec :

$$Z = \begin{pmatrix} Z^{(1)} \\ \\ Z^{(2)} \end{pmatrix}$$

et :

$$A_1 = \varrho V \begin{pmatrix} -\dfrac{\Sigma_{11}}{tr\,(\Sigma_{11}^2)} & \dfrac{\Sigma_{12}}{tr\,(\Sigma_{12}\Sigma_{21})} \\ \dfrac{\Sigma_{21}}{tr\,(\Sigma_{12}\Sigma_{21})} - & \dfrac{\Sigma_{22}}{tr\,(\Sigma_{22}^2)} \end{pmatrix} \tag{2.4}$$

Les mêmes auteurs démontrent que cette fonction d'influence possède les propriétés suivantes :

i) $E[I(X;\varrho V)] = O$

ii) Si $p = q = 1$ alors

$I(X;\varrho V) = I\,(X;\varrho^2) = 2\varrho\left[\frac{-\varrho}{2}\left(\tilde{Y}_1^2 + \tilde{Y}_2^2\right) + \tilde{Y}_1\tilde{Y}_2\right]$ où

$\tilde{Y}_1 = \frac{X_1 - \mu_1}{\tau_1}$ et $\tilde{Y}_2 = \frac{X_2 - \mu_2}{\tau_2}$, τ_1^2 et τ_2^2 désignant respectivement les variances de X_1 et X_2. La partie entre crochets est la fonction d'influence du coefficient de corrélation simple puisque $I(X;\varrho^2) = 2\varrho I(X;\varrho)$.

iii) Si X suit une loi normale $N(\mu,\Sigma)$ alors Var $[I(X;\varrho V)] = \sigma_1^2$ où $\sigma_1^2 = 2tr\,(A_1\Sigma A_1\Sigma)$

iv) Si X suit une loi normale $N(\mu,\Sigma)$ alors $I(X;\varrho V)$ suit la même loi que $\sum_{i=1}^{p+q}\lambda_i W_i^2$ où les W_i sont des variables aléatoires indépendantes $N(O,1)$ et les λ_i sont les valeurs propres de ΣA_1.

A partir de la propriété iv) il est possible de déterminer la loi du maximum des $I(Xj;\varrho V)(X_1,...,X_n$ représentant un échantillon) et du minimum des $I(Xj;\varrho V)$. Ainsi un test peut être effectué pour détecter au maximun deux données aberrantes (celle correspondant au maximum et celle correspondant au minimum, voir Cléroux et alt. (1986)).

En pratique on remplacera les paramètres $\mu^{(i)}$ et $\sum_{ij}$ dans la fonction d'influence par des estimateurs $\hat{\mu}^{(i)}$ et $\widehat{\sum_{ij}}$. Etant donné que l'estimateur habituel de $\sum$, à savoir $S = \frac{1}{n-1}\sum_{j=1}^{n}\left(X_j - \bar{X}\right)\left(X_j - \bar{X}\right)'$, est un mauvais estimateur en présence de données aberrantes, on utilise des estimateurs robustes. Ces estimateurs sont ceux proposés par Huber (1977) et ils sont obtenus par le programme ROBETH développé par Marazzi (1985). On obtient ainsi les estimateurs $\hat{I}\,(X_j;\varrho V)$ des

influences de tout point X_j sur ϱV. D'autre part la procédure de test étant compliquée, il est préférable d'utiliser une procédure heuristique qui consiste à soupçonner comme données douteuses tous les points X_i tels que $\left|\hat{I}(X_j; \varrho V)\right| \geq 3\hat{\sigma}_1$ où $\hat{\sigma}_1$ est obtenu à partir de σ_1 en remplaçant les paramètres inconnus par leurs estimateurs robustes. Les données artificielles de la figure 2.1 montrent à l'évidence que cette procédure n'est pas suffisante.

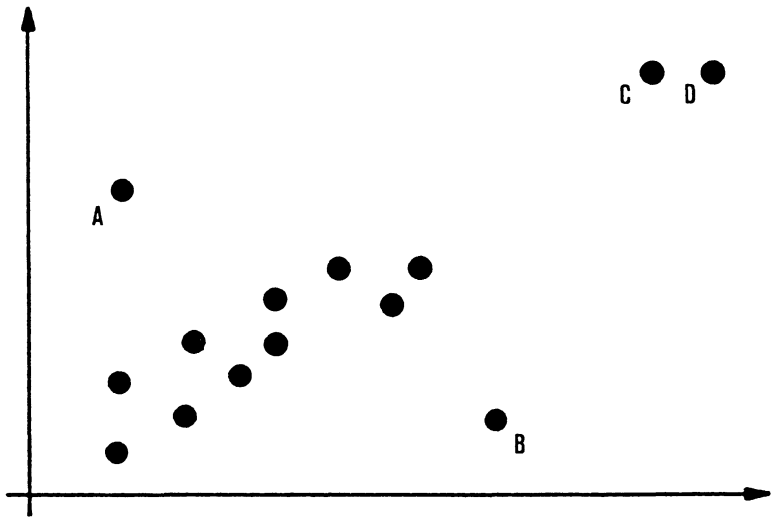


FIGURE 2.1.
Données artificielles dans $\mathbb{R}^2$

A respectivement B auront chacun une influence très forte tandis que C et D auront une influence respective faible. Toutefois le groupe $\{C, D\}$ peut bien être un ensemble aberrant. Ainsi l'heuristique proposée détectera les données aberrantes isolées mais pas nécessairement celles qui forment un groupe. Le paragraphe suivant déterminera comment mesurer l'influence d'un groupe de points et comment former les groupes susceptibles d'être des ensembles douteux.

2.3. Influence d'un groupe de points

2.3.1. Influence de la moyenne d'un groupe de points

En se référant à la figure 2.1, nous avions souligné que les points C et D n'étaient pas détectés comme aberrants. Toutefois si l'on remplace les points C et D par la moyenne de ceux-ci, il est évident que ce point moyen aura une influence très forte. Il est dès lors important de pouvoir calculer l'influence d'un groupe de k points sur le coefficient ρV.

Supposons que nous cherchons l'influence de $Y = \frac{\sum_{i=1}^{k} X_i}{k}$ où les X_i sont des variables aléatoires indépendantes de moyenne μ et de matrice de variance-covariance $\sum$. Il est clair que Y a pour moyenne μ et pour variance-covariance $\frac{\sum}{k}$.

Par application de la formule 2.2 à Y on obtient :

$$I(Y;\rho V) = k\rho V \left[\frac{2U^{(1)'} \sum_{12} U^{(1)}}{tr\left(\sum_{12} \sum_{21}\right)} - \frac{U^{(1)'} \sum_{11} U^{(1)}}{tr\left(\sum_{11}^2\right)} - \frac{U^{(2)'} \sum_{22} U^{(2)}}{tr\left(\sum_{22}^2\right)} \right]$$
$$= kU'A_1U \qquad (2.5)$$

avec $U = \begin{pmatrix} U^{(1)} \\ U^{(2)} \end{pmatrix} = \begin{pmatrix} Y^{(1)} - \mu^{(1)} \\ Y^{(2)} - \mu^{(2)} \end{pmatrix}$ et A_1 est donnée en (2.4).

Le calcul à effectuer est donc très simple puisqu'il suffit de remplacer dans (2.2) X par $Y = \frac{\sum_{i=1}^{k} X_i}{(2.5)k}$ et de multiplier le résultat par k.

La fonction d'influence $I(Y;\varrho V)$ possède les mêmes propriétés que celle de X. Notons que si les $X_i$suivent des lois normales $N(\mu, \sum)$ indépendantes alors

$$\text{Var}[I(Y;\varrho V)] = \text{Var}(kU'A_1U) = k^2 \text{Var}(U'A_1U) = k^2 2tr\left(A_1 \frac{\sum}{k} A_1 \frac{\sum}{k}\right)$$

puisque la matrice de variance-covariance de U est $\frac{\sum}{k}$. Ainsi, $\text{Var}[I(Y;\varrho V)] = \text{Var}[I(X,\varrho V)] = \sigma_1^2$.

Poursuivant l'heuristique proposée en 2.2., il suffit donc de mesurer l'influence d'un groupe quelconque G formé de k points en calculant l'influence échantillonnale $\hat{I}\left(\bar{X}_G;\varrho V\right)$ où $\bar{X}_G$ = moyenne des points X_i du groupe G. Le groupe est soupçonné douteux si $\left|\hat{I}\left(\bar{X}_G;\varrho V\right)\right| \geq 3\hat{\sigma}_1$ comme dans le cas d'un point. Le problème est alors de former des groupes sans devoir si possible énumérer tous les doublets, triplets etc...

2.3.2. Formation de groupes pertinents

Pour former un groupe de données douteuses, il est nécessaire que l'ensemble des points de ce groupe constitue un amas compact de points. Il est alors naturel d'utiliser pour ce faire une procédure de classification hiérarchique. Les groupes pertinents sont ceux dont les éléments s'unissent à un bas niveau de distance dans

la classification et restent ensemble longtemps avant de s'associer à d'autres points. Pour mesurer l'influence des groupes pertinents, on utilise la procédure développée en 2.3.1.

2.4. Un exemple

Considérons les données d'une étude sur le rendement de vendeurs par rapport à leurs capacités intellectuelles. Les données ont été recueillies par Johnson et Wichern (1982). Il s'agit de 50 observations des variables suivantes :

$$X^{(1)} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = \begin{pmatrix} \text{Croissance des ventes} \\ \text{Profit provenant des ventes} \\ \text{Ventes relatives à de nouveaux clients} \end{pmatrix}$$

$$X^{(2)} = \begin{pmatrix} X_4 \\ X_5 \\ X_6 \\ X_7 \end{pmatrix} = \begin{pmatrix} \text{Note d'esprit de créativité} \\ \text{Note d'habileté mécanique} \\ \text{Note d'esprit d'abstraction} \\ \text{Note d'esprit mathématique} \end{pmatrix}$$

La procédure proposée en 2.2. avec les estimations robustes laisse apparaître le tableau suivant des 4 plus fortes influences individuelles :

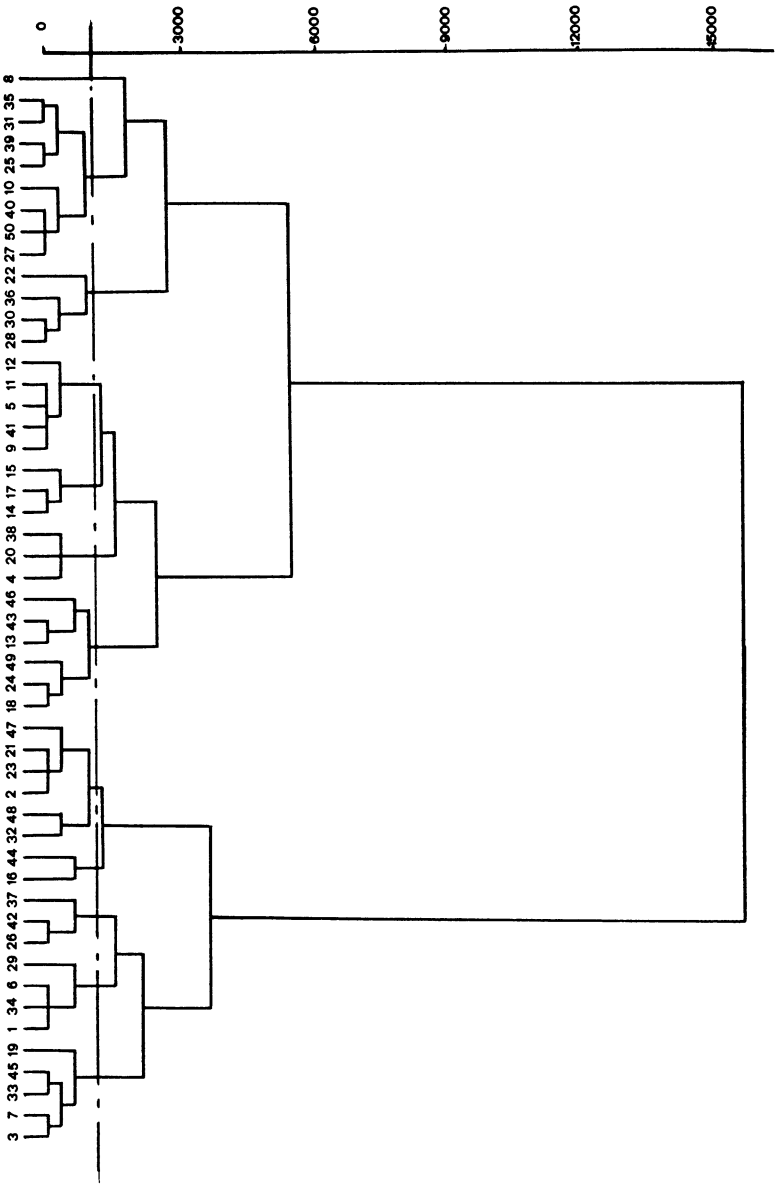
Point	36	22	28	37
Influence négative	−0.374	−0.309	−0.272	−0.192

Point	23	2	44	48
Influence positive	0,149	0,173	0,221	0,235

La valeur du coefficient $R\mathrm{V}^{2}$ robuste entre $X^{(1)}$ et $X^{(2)}$ (i.e. coefficient calculé à partir d'une estimation robuste de$\sum$, cf. § 2.2) est 0,9469 et celle de $\hat{\sigma}_1$ est 0,1176. Ainsi le seul point douteux au sens de l'heuristique est le point 36.

Deux autres méthodes de détection de données aberrantes ont été utilisées sur ces données : la première due à Rohlf (1975) est basée sur la recherche des arêtes de longueur maximale dans un arbre minimum et la seconde due à Wilks (1963) calcule le rapport des déterminants de matrices de données. Les deux procédures donnent lieu à un test statistique en supposant la normalité des données. Les points 8 et 44 sont ceux qui fournissent les plus longues arêtes par la procédure de Rohlf, c'est-à-dire ceux qui sont les plus douteux. Par la méthode de Wilks, ce sont les points 8 et 10 qui fournissent les rapports les plus faibles des déterminants et par conséquent ce sont les points les plus douteux. Toutefois, dans les deux cas, les tests de détection de données aberrantes ne sont pas significatifs à 5 %.

Utilisant la procédure CLUSTER (mesure SEUCLID et méthode de WARD) de SPSS X, on obtient le dendrogramme suivant :



En formant les groupes fournis par la coupure indiquée par un trait d'axe, on obtient en ne retenant que ceux de cardinalité inférieure ou égale à 5 (10 % des données) :

Groupe	Influence $(3\hat{\sigma}_1 = 0.3529)$
$G_1 = \{8\}$	0.1451
$G_2 = \{28, 30, 36, 22\}$	−0.7762
$G_3 = \{4, 20, 38\}$	−0.0373
$G_4 = \{16, 44\}$	0.3803
$G_5 = \{26, 42, 37\}$	−0.1196
$G_6 = \{1, 34, 6, 29\}$	0.2654
$G_7 = \{3, 7, 33, 45, 19\}$	−0.0196
$G_8 = \{14, 15, 17\}$	−0.2157
$G_9 = \{9, 41, 5, 11, 12\}$	−0.0390

En appliquant la procédure de § 2.3, on peut avoir des soupçons sérieux sur les groupes $G_2 = \{28, 30, 36, 22\}$ et $G_4 = \{16, 44\}$. Remarquons que 36 seul était déjà douteux mais qu'en définitive il forme un groupe douteux en compagnie des points 28, 30 et 22.

3. Le coefficient $\varrho V reg$: Fonction d'influence et données aberrantes

3.1. Influence d'un point et d'un groupe de points

Dans le contexte plus particulier de la régression linéaire multivariée, les développements effectués dans le paragraphe 2 peuvent aussi se faire. Robert et Escoufier (1976) définissent le coefficient $\varrho V reg$ comme le maximum de $\varrho V\left(X^{(1)}, MX^{(2)}\right)$ sur toutes les matrices M telles que $\sum_{12} M - M' \sum_{22} M = 0$. On a alors :

$$\varrho V reg\left(X^{(1)}, X^{(2)}\right) = \left[\frac{tr\left(\sum_{11} - \sum_{11.2}\right)^2}{tr \sum_{11}^2}\right]^{1/2} = \left[\frac{tr\left(\sum_{12} \sum_{22}^{-1} \sum_{21}\right)^2}{tr \sum_{11}^2}\right]^{1/2} \tag{3.1}$$

où :

$$\sum_{11.2} = \sum_{11} - \sum_{12} \sum_{22}^{-1} \sum_{21}$$

Si $p = 1$, $\varrho V_{reg} = R^2$, le carré du coefficient de corrélation multiple de $X^{(1)}$ par rapport à $X^{(2)}$. Cléroux et al (1986) ont obtenu la fonction d'influence $I(X; \varrho V_{reg})$ d'un point X sur le coefficient ϱV_{reg} :

$$I(X;\varrho V_{reg}) = \varrho V_{reg}\left(\frac{2Z^{(1)'}\sum_{11}^{*}BZ^{(2)}}{tr\left(\sum_{11}^{*2}\right)} - \frac{Z^{(1)'}\sum_{11}Z^{(1)}}{tr\left(\sum_{11}^{2}\right)} - \frac{Z^{(2)'}B'\sum_{11}^{*}BZ^{(2)}}{tr\left(\sum_{11}^{*2}\right)}\right) \tag{3.2}$$

où $Z^{(i)} = X^{(i)} - \mu^{(i)}, (i = 1,2), \sum_{11}^{*} = \sum_{11} - \sum_{11.2} = \sum_{12}\sum_{22}^{-1}\sum_{21}$ et $B = \sum_{12}\sum_{22}^{-1}$ la matrice des coefficients de régression de $X^{(1)}$ sur $X^{(2)}$. Comme pour celle de ϱV cette fonction d'influence peut se mettre sous la forme d'une forme quadratique :

$$I(X;\varrho V_{reg}) = Z'A_2Z \tag{3.3}$$

avec :

$$Z = \begin{pmatrix} Z^{(1)} \\ \\ Z^{(2)} \end{pmatrix}$$

et :

$$A_2 = \varrho V_{reg}\begin{pmatrix} -\dfrac{\Sigma_{11}}{tr\left(\Sigma_{11}^{2}\right)} & \dfrac{\Sigma_{11}^{*}B}{tr\left(\Sigma_{11}^{*2}\right)} \\ \dfrac{B'\Sigma_{11}^{*}}{tr\left(\Sigma_{11}^{*2}\right)} & -\dfrac{B'\Sigma_{11}^{*}B}{tr\left(\Sigma_{11}^{*2}\right)} \end{pmatrix} \tag{3.4}$$

Dans le même article, les auteurs montrent les propriétés suivantes :

i) $E(I(X;\varrho V_{reg})) = 0$

ii) Si $p = 1$ alors
$I(X;\varrho V_{reg}) = \frac{1}{\sigma_1^2}\left[2Z_1\beta'Z^{(2)} - Z_1^2R^2 - \left(\beta'Z^{(2)}\right)^2\right]$ où
$\beta = \sum_{22}^{-1}\sigma_{(1)}$ représente le vecteur des coefficients de la régression de $X^{(1)} = X_1$ sur $X^{(2)}$ et R le coefficient de corrélation multiple entre X_1 et $X^{(2)}$, $\sigma_{(1)}$ étant le vecteur colonne $\sum_{21}$. Ce cas particulier est traité dans Cléroux et alt. (1989).

iii) Si X suit une loi normale $N(\mu, \sum)$ alors
$Var[I(X;\varrho V_{reg})] = \sigma_2^2$ où $\sigma_2^2 = 2tr(A_2 \sum A_2 \sum)$

iv) Si X suit une loi normale $N(\mu, \sum)$ alors $I(X;\varrho V_{reg})$ suit la loi de $\sum_{i=1}^{p+q}\lambda_i W_i^2$ où les W_i sont des variables $N(O,1)$ indépendantes et les λ_i sont les valeurs propres de $\sum A_2$.

En présence d'un échantillon, on remplacera dans l'expression de $\varrho Vreg$ les paramètres par les estimateurs habituels. Ainsi on obtient RV_{reg} l'estimateur de ϱV_{reg}.

Comme dans le cas de la fonction d'influence $I(X;\varrho V)$ il est possible de développer à partir de iv) une procédure de test. On lui préfère l'heuristique développée au § 2.2 en utilisant les mêmes estimateurs robustes pour estimer les paramètres de $I(X;\varrho V_{reg})$. Ainsi, en présence d'un échantillon $X_1,...,X_n$ on obtiendra les valeurs échantillonnales $\hat{I}(X_j;\varrho V_{reg})$ que l'on comparera à $\pm 3\hat{\sigma}_2$.

Pour ce qui est des influences de groupes de points, il est possible de procéder de façon semblable au § 2.3 en utilisant la même procédure de classification hiérarchique. En effet, ni la définition de l'influence d'un groupe de points, ni le principe de regroupement des points ne sont influencés par le fait que l'on travaille dans le cadre d'un modèle de régression multivariée.

3.2. Exemple

Woltz et al (1948) ont analysé un échantillon de 25 feuilles de tabac en fonction de constituants organiques et inorganiques. Sur chacune des feuilles, ils ont relevé les variables suivantes :

$$X^{(1)} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = \begin{bmatrix} \text{Taux de consommation en pouces par 1000 secondes} \\ \text{Pourcentage de sucre} \\ \text{Pourcentage de nicotine} \end{bmatrix}$$

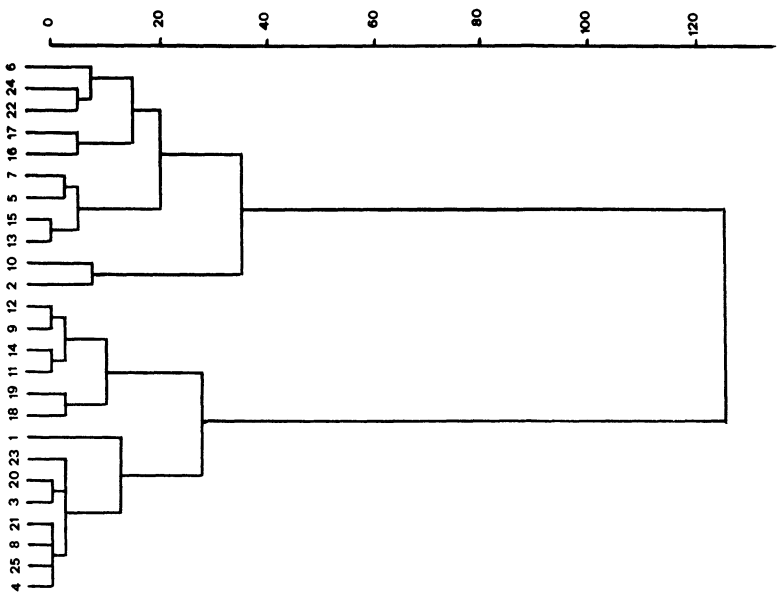
$$X^{(2)} = \begin{bmatrix} X_4 \\ X_5 \\ X_6 \\ X_7 \\ X_8 \\ X_9 \end{bmatrix} = \begin{bmatrix} \text{Pourcentage de nitrogène} \\ \text{Pourcentage de chlore} \\ \text{Pourcentage de potassium} \\ \text{Pourcentage de phosphore} \\ \text{Pourcentage de calcium} \\ \text{Pourcentage de magnésium} \end{bmatrix}$$

Les influences ordonnées donnent le tableau suivant :

Point X_i	22	14	36	12	32	1	24	17	18	19	4
$\hat{I}(X_i;\varrho V_{reg})$	−1.31	−.56	−.45	−.30	−.29	−.21	−.09	−.07	−.05	−.02	−.01

Point X_i	16	8	11	9	25	20	21	13	15	23	7	5	2	10
$\hat{I}(X_i;\varrho V_{reg})$	.01	.05	.07	.08	.15	.19	.20	.27	.27	.28	.30	.42	.61	.67

La valeur du coefficient RV_{reg} robuste est 0,7421 et celle de $\hat{\sigma}_2$ est 0.4439. Aucun point n'est donc aberrant au sens de l'heuristique proposée. A l'aide de la méthode de Rohlf, on constate que le groupe $\{2,10\}$ est détecté comme aberrant $(\alpha = 0,05)$. Avec la méthode de Wilks, la conclusion est identique. La classification a donné le dendrogramme suivant :



Les groupes intéressants sont $G_1 = \{2, 10\}$ et $G_2 = \{16, 17\}$dont les influences respectives sont 1.62 et O.23. On constate alors que G_1est un groupe douteux au sens de l'heuristique proposée comme pour les deux autres méthodes.

Bibliographie

ANDREWS, D.F. (1972) - Plots of High-dimensional Data, *Biometrics, 28, 125-136.*

ANDREWS, D.F. et PREGIBON, D. (1978) - Finding the Outliers that Matter, *Journal of Royal Statistical Society, Series B, 40, N°1, 85-93.*

BACON-SHONE, J. et FUNG, W.K. (1987) - A New Graphical Method for Detecting Single and Multiple Outliers in Univariate and Multivariate Data, *Appl. Statist.*, 36, 153-162.

BARNETT, V. (1979) - Some Outlier Tests for Multivariate Samples, *South African Statistical Journal, 13, 29-52.*

BARNETT, V. (1983) - Reduced Distance Measures and Transformations in Processing Multivariate Outliers, *Australian Journal of Statistics, 25* (1), 64-75.

BECKMAN, R.J. et COOK, R.D. (1983) - Outlier..s, *Technometrics, 25, 119-163.*

BRADU, D. et HAWKINS, D.M. (1982) - Location of Multiple Outliers in Two-way Tables Using Tetrads, *Technometrics, 24, 103-108.*

CAMPBELL, N.A. (1978) - The Influence Function as an Aid in Outlier Detection in Discriminant Analysis, *Applied Statistics*, 27, 251-258

CHATTERJEE, S. et HADI, A.S. (1986) - Influential Observations, High Leverage Points, and Outliers in Linear Regression, *Statistical Science, 1, 379-416.*

CHERNOFF, H. (1973) - Using Faces to Represent Points in k-Dimensional Space Graphically, *Journal of the American Statistical Association, 68, 361-368.*

CHERNOFF, H. et RIZVI, M.H (1975) - Effect on Classification Error of Random Permutations of Features in Representing Multivariate Data by Faces, *Journal of the American Statistical Association, 70, 548-554.*

CLEROUX, R., HELBLING, J.M. et RANGER, N. (1986) - Some Methods of Detecting Multivariate Outliers, *Computational Statistics Quartely, 3, 177-195.*

CLEROUX, R., HELBLING, J.M. et RANGER, N. (1989), Influential Subsets Diagnostics Based on Multiple Correlation, *Computational Statistics Quartely,* 2, 99-117.

COOK, R.D. (1977), Detection of Influential Observations in Linear Regression, *Technometrics,* 19, 15-18.

COOK, R.D. (1979) - Influential Observations in Linear Regression, *Journal of the American Statistical Association, 74, 169-174.*

COOK, R.D. et WEISBERG, S. (1980) - Characterizations of an Empirical Influence Function for Detecting Influential Cases in Regression, *Technometrics, 22, 495-508.*

DAS, R. et SINHA, B.K. (1986) - Detection of Multivariate Outliers with Dispersion Slippage in Elliptically Symetric Distributions, *Annals of Statist., 14, 1619-1624.*

DEVLIN, S.J., GNANADESIKAN, R. et KETTENRING, J.R. (1975) - Robust Estimation and Outlier Detection with Correlation Coefficients, *Biometrika, 62, 531-545.*

DRAPER, N.R., GUTTMAN, I. et JOHN, J.A. (1984) - Premium and Protection of a Response Estimation Procedure for Two-Way Tables when Outliers Occur, *Computational Statistics and Data Analysis, 2,229-236.*

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

GAREL, B. (1978) - Tests de détection de valeurs aberrantes multidimensionnelles, *Annales de l'institut Henri Poincarré, 14, N° 3, 303-314.*

GENTLEMAN, J.F. (1980), Some Methods of Searching for Outliers, Multivariate Statistical Analysis, R.P. Gupta (ed.), New York, *North Holland.*

GNANADESIKAN, R. (1977), Methods for Statistical Data Analysis of Multivariate Observations, New York, *John Wiley and Sons.*

GNANADESIKAN, R. et KETTENRING, J.R. (1972) - Robust Estimates, Residuals, and Outlier Detection with Multiresponse Data, *Biometrics, 28, 81-124.*

GRUBBS, F.E. (1969), Procedures for Detecting Outlying Observations in Samples, *Technometrics, 11, 1-21.*

GUTTMAN, I. (1973) - Care and Handling of Univariate or Multivariate Outliers in Detecting Spuriosity - A Bayesian Approach, *Technometrics, 15, 723-738*

HAMPEL, F.R. (1974) - The influence Curve and its Role in Robust Estimation,

Journal of the American Statistical Association, 69, 383-393.

HAMPEL, F.R. (1985) - The Breakdown Points of the Mean Combined with some Rejection Rules, *Technometrics, 27, 95-107.*

HAWKINS, D.M. (1974) - The Detection of Errors in Multivariate Data Using Principal Components, *Journal of the American Statistical Association, 69, 340-344.*

HAWKINS, D.M. (1980) - Identification of Outliers, London, *Chapman and Hall.*

HAWKINS, D.M., BRADU, D. et KASS, G.V. (1984), Location of Several Outliers in Multiple-Regression Data Using Elemental Sets, *Technometrics*, 26, 197-208.

HUBER, P.J. (1977), Robust Covariances, in Statistical Decision Theory and Related Topics, Vol. 2, ed. by S.S. Grupha and D.S. Moore, *Academic Press, 165-191.*

JAIN, R.B. (1981) - Detecting Outliers : Power and Some Other Considerations, *Communications in Statistics, 10, 2299-2314.*

JOHNSON, R.A. et WICHERN, D.W. (1982) - Applied Multivariate Statistical Analysis, Englewood Cliffs N.J., *Prentice Hall.*

KALE, B.K. (1976) - Detection of Outliers, *Sankhya, Series B, 38, 356-363.*

KALE, B.K (1979) - Outliers - A review, *Journal of the Indian Statistical Association, 17, 51-67.*

KOTZE, T.J.V. et HAWKINS, D.M. (1984) - The Identification of Outliers in Two-Way Contingency Tables Using 2x2 Subtables, *Applied Statistics, 33, 215-223.*

MARAZZI, A. (1985) - Robust Affine Invariant Covariances, Doc N°6, Division de Statistique et Informatique, *Institut Universitaire de Médecine Sociale et Préventive, Lausanne.*

McCULLOCH, C.E. et MEETER, D. (1983), Discussion of "Outlier..s" by Beckman, R.J. and Cook, R.D., *Technometrics, 25, 152-155.*

PRESCOTT, P. (1980) - A Review of Some Robust Data Analysis and Multiple Outlier Detection Procedures, *Bulletin in Applied Statistics, 7, 141-158.*

ROBERT, P. et ESCOUFIER, Y. (1976) - A Unifying Tool for Linear Multivariate Statistical Methods : The RV-Coefficient, *Applied Statistics, 25, 257-265.*

ROHLF, F.J. (1975) - Generalization of the Gap Test for the Detection of Multivariate Outliers, *Biometrics, 31, 93-101.*

ROSNER, B. (1983) - Percentage Points of a Generalized ESD Many Outlier Detection, *Technometrics, 25, 165-172.*

ROUSSEEUW, P.J. et LEROY, A.M. (1987) - Robust Regression and Outlier Detection, *John Wiley and Sons.*

ROUSSEEUW, P.J. et Van ZOMEREN, B.C (1987) - Identification of Multivariate Outliers and Leverage Points by Means of Robust Covariance Matrices, *Report 87-15 of the Faculty of Mathematics and Informatics, Delft University of Technology.*

SCHALL, R. et DUNNE, T.T. (1987) - On Outliers and Influence in the General Multivariate Normal Linear Model, *Proc. Second International Tampere Conference in Statistics.*

SCHWAGER, S.J. et MARGOLIN, B.H. (1982) - Detection of Multivariate Normal Outliers, *The Annals of Statistics*, 10, 943-954.

SINHA, B.K. (1984), Detection of Multivariate Outliers in Elliptically Symmetric Distributions, *The Annals of Statistics, 12, 1558-1565.*

SIOTANI, M. (1959) - The Extreme Value of the Generalized Distances of the Individual Points in the Multivariate Sample, *Annals of the Institute of Statistical Mathematics, 10, 183-203.*

TEILLARD, P. et VOLLE, M. (1976) - Détection de points aberrants en analyse factorielle des correspondances, *Annales de l'INSEE, 22-23, 237-253.*

WILKS, S.S. (1963) - Multivariate Statistical Outliers, *Sankhya*, 25, 407-426.

WOLTZ, W.G., REID, W.A. et COLWELL, W.E. (1948) - Sugar and Nicotine in Cured Bright Tabacco as Related to Mineral Element Composition, *Proc. Soil Sci. Am., 13, 385-387.*