REVUE DE STATISTIQUE APPLIQUÉE

LOUIS-PAUL RIVEST
NATHALIE PLANTE

L'analyse en composantes principales robuste

Revue de statistique appliquée, tome 36, n° 1 (1988), p. 55-66
<http://www.numdam.org/item?id=RSA_1988__36_1_55_0>

Rev. Statistique Appliquée, 1988, **XXXVI** (1), 55-66

L'ANALYSE EN COMPOSANTES PRINCIPALES ROBUSTE

Louis-Paul RIVEST et Nathalie PLANTE

Département de mathématiques, statistique et actuariat
Université Laval, Cité Universitaire, Québec, G1K 7P4, Canada

RÉSUMÉ

On étudie des analyses en composantes principales où chaque individu reçoit un poids déterminé au préalable lors d'une estimation robuste des paramètres de localisation et de dispersion de l'échantillon des individus. La classe d'estimateurs de MARONNA, qui constitue une généralisation multivariée des M-estimateurs est utilisée à cette fin. Deux procédures robustes sont considérées : celle de HUBER et le biweight de TUKEY. Cette dernière donne un poids nul aux individus jugés aberrants. Deux exemples sont présentés : l'étude d'un échantillon simulé et une analyse robuste des données de JAMBU (voir [11]).

Mots clés : *Analyse en Composantes Principales, Biweight, Estimateur de Huber, Méthodes Robustes, Poids d'une unité statistique.*

ABSTRACT

This work is concerned with principal components analysis where each individual is given a weight, determined a priori in a robust estimation of the location and scale parameters of the sample of individuals. MARONNA's multivariate generalization of M-estimators is used for this purpose. Two robust estimators are considered : the one of HUBER, and TUKEY's biweight which give a zero weight to outliers. Two examples are presented : the study of a simulated sample and a robust analysis of JAMBU's data (see [11]).

1. Introduction

L'influence indue que quelques observations, douteuses ou aberrantes, peuvent avoir sur le résultat d'une analyse statistique préoccupe les statisticiens depuis fort longtemps ([8]). Certains auteurs se sont penchés sur ce problème en analyse en composantes principales (A.C.P.). Ainsi VOLLE, dans [11] p. 91, utilise les aides à l'interprétation pour repérer les points aberrants. Il suggère de refaire l'A.C.P. sans les points aberrants et de mettre ces derniers en points supplémentaires. BENASSENI, dans [1], étudie l'impact que des données douteuses ont sur les valeurs propres. CRITCHLEY, quant à lui, suggère dans [2] quelques « mesures d'influence » pour détecter les points aberrants.

LEBART, MORINEAU et TABART proposent, dans [6], en présence de données « hétérogènes », de faire l'A.C.P. sur les rangs, ou une fonction des rangs des observations. Cette approche a l'avantage de la simplicité. Cependant, en s'éloignant des données originales on obtient des résultats qui sont souvent difficiles à interpréter.

Ce travail étudie une troisième voie : les A.C.P. liées à des estimateurs robustes des paramètres de localisation (μ) et de dispersion (Σ) d'un échantillon multivarié. Si $\{x_i'\}$ représente les lignes du fichier de données, les estimateurs $\hat{\mu}$ et $\hat{\Sigma}$ sont définis implicitement par

$$\hat{\mu} = \frac{\Sigma\, m_i^{1/2}\, x_i}{\Sigma\, m_i^{1/2}} \tag{1}$$

$$\hat{\Sigma} = \frac{\Sigma\, m_i\,(x_i - \hat{\mu})\,(x_i - \hat{\mu})'}{(n-1)\,CT} \tag{2}$$

où $m_i = g((x_i - \hat{\mu})'\,\hat{\Sigma}^{-1}(x_i - \hat{\mu}))$ est une fonction, en général décroissante, de $(x_i - \hat{\mu})'\,\hat{\Sigma}^{-1}(x_i - \hat{\mu})$, la distance de Mahalanobis entre x_i et $\hat{\mu}$ et

$$CT = E(Wg(W))/k, \tag{3}$$

W étant une variable aléatoire χ^2 à k degrés de liberté (k est le nombre de variables). Une fois le calcul des estimations robustes de μ et Σ terminé, à chaque individu x_i, on associe le poids $P_i = m_i/((n-1)\,CT)$. Dans ce travail, on étudie les A.C.P. du triplet (Y,P,I) où P est la matrice diagonale $n \times n$ des P_i, Y est le tableau des données centrées et réduites ($y_i = \hat{D}^{-1/2}(x_i - \hat{\mu})$ où $\hat{D}$ est la matrice diagonale $k \times k$ des $\hat{\Sigma}_{jj}$).

2. Estimation robuste de μ et Σ

Supposons, pour l'instant, que $\{x_i\}$ constitue un échantillon d'une distribution F(x) définie dans R^k. MARONNA [7], HUBER [5] et HAMPEL, RONCHETTI, ROUSSEEUW et STAHEL ([4]) suggèrent d'utiliser les estimateurs définis implicitement par (1) et (2) lorsque F appartient à un voisinage d'une loi normale. Sous certaines hypothèses de symétrie, ils ont obtenu la distribution asymptotique, lorsque n devient grand, de $\hat{\mu}$ et $\hat{\Sigma}$ (voir aussi [9]). Ils ont, en outre, montré que l'estimateur de HUBER, obtenu avec $g_H(x) = \min(1, c/x)$, pour un c donné, est asymptotiquement optimal : il est « minimax » pour Σ ([5], p. 233) et « B-robuste » pour μ et une mesure de la « forme » de Σ ([4], p. 293-294).

Les paramètres estimés par la solution de (1) et (2) sont définis implicitement par

$$E_F(g^{1/2}((x-\mu)'\,\Sigma^{-1}(x-\mu))\,(x-\mu)) = 0 \tag{4}$$

$$E_F\left(\frac{g((x-\mu)'\,\Sigma^{-1}(x-\mu))\,(x-\mu)\,(x-\mu)'}{CT}\right) = \Sigma. \tag{5}$$

HUBER dans [5], p. 215-223, montre, sous des conditions de régularité restrictives, que ces équations ont une solution unique.

Si F est une loi normale multivariée de moyenne μ_0 et de matrice de variances Σ_0, notre choix de CT implique que μ_0 et Σ_0 sont solutions de (4) et (5) (voir [7], p. 53, exemple 1 pour le cas de l'estimateur de HUBER).

En plus des estimateurs de HUBER, nous allons également utiliser la classe d'estimateurs liée à la fonction biweight de TUKEY. Ils utilisent la fonction de poids

$$g_T(x) = \begin{cases} \left(1 - \frac{x}{c}\right)^4 & \text{si } x < c \\ 0 & \text{si } x > c \end{cases}$$

où c est une constante arbitraire.

2.1. *Algorithme pour le calcul de $\hat{\mu}$ et $\hat{\Sigma}$*

En général, (1) et (2) ont, pour l'estimateur de HUBER, une solution unique. Cependant, ce résultat n'a pas encore été démontré formellement sous des conditions de régularité acceptables (voir [5] et [10]). Pour calculer les estimations de HUBER, nous utiliserons l'algorithme suggéré dans [5], p. 238.

Etant donné A une matrice définie positive, soit $A^{1/2}$ la matrice triangulaire inférieure ayant des éléments positifs sur la diagonale et satisfaisant $A = A^{1/2}(A^{1/2})'$ (le produit $A^{1/2}(A^{1/2})'$ est appelé la décomposition de Choleski de A). Soit $\hat{V} = (\hat{\Sigma}^{1/2})^{-1} = \hat{\Sigma}^{-1/2}$. Les valeurs de départ de l'algorithme sont, si $\overline{x}$ désigne la moyenne usuelle des x_i :

$$t_0 = \overline{x},\ V_0 = \left[\frac{\Sigma (x_i - \overline{x})\,(x_i - \overline{x})'}{n - 1}\right]^{-1/2}$$

Chaque itération se compose de deux étapes : l'étape échelle et l'étape localisation. Soient t_j et V_j les valeurs des paramètres obtenues après j itérations; t_{j+1} et V_{j+1} sont calculées de la manière suivante :

Etape échelle

Soient $w_i = V_j(x_i - t_j)$ et

$$C_{j+1} = \frac{\sum_i g_H(w_i' w_i)\, w_i w_i'}{(n - 1)\, CT}$$

où, d'après (3),

$$CT = \frac{\int_0^\infty \min(x, c)\, f_k(x)\, dx}{k}, \qquad (6)$$

c est la constante apparaissant dans la fonction g_H et $f_k(x)$ est la densité d'une χ^2 à k degrés de liberté. Alors $V_{j+1} = C_{j+1}^{-1/2} V_j$.

Etape localisation

Soient $w_i = V_{j+1}(x_i - t_j)$ et

$$h_{j+1} = \frac{\Sigma\, g_H^{1/2}(w_i' w_i)\,(x_i - t_j)}{\Sigma\, g_H^{1/2}(w_i' w_i)}$$

alors prendre $t_{j+1} = t_j + h_{j+1}$.

« Le critère de convergence est donné par : arrêter si la plus grande composante, en valeur absolue, de h_j et de $(C_j - I)$ est inférieure à un ε prédéterminé » (en général $\varepsilon = 10^{-3}$). Quand un j satisfaisant ce critère est atteint, on prend $\hat{\Sigma} = V_j^{-1}(V_j^{-1})'$, $\hat{\mu} = t_j$.

Cet algorithme a les propriétés suivantes :

— Quand k = 1, c'est l'algorithme de HUBER-DUTTER pour le calcul des estimateurs des paramètres de localisation et d'échelle appelé « Proposal 2 » de HUBER. Il converge ([5], section 7.8).

— Quand k > 1, l'algorithme contenant seulement des étapes localisation pour l'estimation de μ par (1) lorsque $\hat{\Sigma}$ est connue converge pour g_H ([5], p. 239). Il en est de même pour l'algorithme contenant seulement des étapes échelle pour l'estimation de Σ par (2) lorsque $\hat{\mu}$ est connu et c est assez grand, ([5], p. 216). Cet algorithme est également étudié par TYLER dans [10].

Ainsi, même s'il n'existe aucune preuve formelle de la convergence de cet algorithme avec la fonction g_H, nous l'avons utilisé sans problème. Pour l'estimateur obtenu avec la fonction g_T, il existe, en général, plusieurs solutions à (1) et (2) (voir [3]); l'algorithme est sensible à la valeur de départ. Ainsi, suivant une suggestion de HUBER, nous nous contenterons d'utiliser une approximation des estimations biweight obtenue après trois itérations de l'algorithme avec la fonction g_T, à partir des estimations de HUBER.

Une fois $\hat{\Sigma}$, l'estimateur robuste de Σ, calculé, on obtient une estimation robuste $\hat{R}$ de la matrice de corrélation par

$$\hat{R} = \hat{D}^{-1/2} \hat{\Sigma} \hat{D}^{-1/2}$$

où $\hat{D}$ est la matrice diagonale, k × k des éléments sur la diagonale de $\hat{\Sigma}$.

3. L'analyse en composantes principales robuste

Notons d'abord que la somme des poids des individus, $\Sigma m_i/((n-1)CT)$, n'est pas égale à un. Ces poids permettent d'obtenir des résultats en tout point comparables à ceux de l'estimation classique : quand les données sont normales, les deux estimateurs de Σ estiment le même paramètre.

3.1. L'analyse dans R^k

Les vecteurs unitaires utilisés pour la représentation des individus sont les vecteurs propres de

$$Y' P Y = \hat{R}$$

Si u_j est le vecteur propre correspondant à la $j^{ème}$ valeur propre, la coordonnée de la projection du $i^{ème}$ individu sur le $j^{ème}$ axe factoriel est $u_j' \hat{D}^{-1/2}(x_i - \hat{\mu})$.

Notons que

$$\hat{R} = \frac{\Sigma z_i z_i'}{n-1}$$

où $z_i = (CT\hat{D})^{-1/2}(x_i^w - \hat{\mu})$ et $x_i^w = \hat{\mu} + m_i^{1/2}(x_i - \hat{\mu})$. Ainsi déterminer les axes principaux des triplets (Y, P, I) et (Z, I/(n − 1), I) sont deux actions équivalentes (Z est le tableau dont la $i^{ème}$ ligne est z_i'). En A.C.P. robuste, on remplace x_i par x_i^w, sa valeur « winsorisée ». En général $m_i \leqslant 1$ et x_i^w se trouve sur le segment joignant x_i et $\hat{\mu}$. Avec l'estimateur de HUBER, pour les x_i situés à l'intérieur de l'ellipsoïde défini par

$$\{ t : (t - \hat{\mu})' \hat{\Sigma}^{-1} (t - \hat{\mu}) = c \}, \tag{7}$$

$x_i = x_i^w$ tandis que si x_i est à l'extérieur de cette ellipsoïde, x_i^w est le point d'intersection entre le segment joignant x_i et $\hat{\mu}$ et la frontière de (7). Cette propriété caractérise les estimateurs de HUBER : ils ramènent les observations « près » de leurs valeurs prédites (ici $\hat{\mu}$) tout en les faisant participer à la procédure d'estimation ([5], p. 19).

L'estimateur biweight accorde un traitement différent aux individus « éloignés ». Si $y_i' \hat{R} y_i$ est plus grand que c, le $i^{ème}$ individu se voit accorder un poids nul et n'est pas utilisé dans le calcul des estimations de μ et Σ. Il constitue, ipso facto, un point supplémentaire. Ainsi l'analyse biweight reflète uniquement la structure des données qui ne sont pas « trop éloignées » de $\hat{\mu}$, contrairement à l'analyse de HUBER qui « conserve » ces observations après les avoir ramenées près de $\hat{\mu}$.

3.2. L'analyse dans R^n

Cette analyse est très semblable à l'analyse classique. Les vecteurs propres v_j utilisés pour la représentation des variables satisfont

$$v_j = P^{1/2} Y u_j / \sqrt{\lambda_j},$$

où λ_j est la $j^{ème}$ valeur propre. La distance entre deux variables j et j' s'écrit

$$d^2(j, j') = 2(1 - \hat{R}_{jj'})$$

où $\hat{R}_{jj'}$ est le coefficient de corrélation robuste entre j et j'. C'est l'élément (j, j') de $\hat{R}$.

3.3. Les aides à l'interprétation

La contribution du $i^{ème}$ individu à λ_j est égale à ([11], p. 118)

$$CN_i = \frac{m_i (u_j' y_i)^2}{(n - 1) CT \lambda_j}$$

Proposition

Pour les A.C.P. utilisant les procédures robustes d'estimation décrites à la section 2,

$$CN_i \leqslant \frac{\max_x x\, g(x)}{(n - 1) CT}$$

avec égalité si et seulement si $y_i = (x_0 \lambda_j)^{1/2} u_j$ où x_0 est une valeur de x pour laquelle $x_0 g(x_0) = \max_x x\, g(x)$.

Démonstration

On peut écrire

$$CN_i = \frac{m_i (u_j' \hat{R}^{1/2} \hat{R}^{-1/2} y_i)^2}{(n - 1) CT \lambda_j}$$

Par l'inégalité de Cauchy Schwarz,

$$\begin{aligned} CN_i &\leqslant \frac{m_i \lambda_j y_i' \hat{R}^{-1} y_i}{(n-1)\,CT\,\lambda_j} \\ &= \frac{g(y_i' \hat{R}^{-1} y_i)\; y_i' \hat{R}^{-1} y_i}{(n-1)\,CT} \\ &\leqslant \max_x \frac{x\, g(x)}{(n-1)\,CT} \end{aligned}$$

avec égalité si et seulement si $y_i = (x_0 \lambda_j)^{1/2} u_j$.

C.Q.F.D.

Ainsi, en A.C.P. robuste, les contributions individuelles aux valeurs propres sont bornées. Pour la méthode de HUBER, la borne est de $c/((n-1)\,CT)$. Cette quantité atteint son minimum, $k/(n-1)$, pour $c = 0$, car, d'après (6), $CT/c \leqslant 1/k$ et CT/c tend vers $1/k$ quand c tend vers 0. Pour l'estimateur biweight la borne est de $.08\ c/((n-1)\,CT)$.

Les autres aides à l'interprétation sont semblables à ceux de l'analyse classique.

4. Deux exemples

Les analyses en composantes principales robuste et classique donnent souvent des résultats semblables. C'est seulement en présence de données hétérogènes que les deux analyses diffèrent. Les deux exemples suivants utilisent de telles données. En effet, il est plus intéressant de présenter des cas où apparaissent des différences entre les deux analyses car ils illustrent mieux les caractéristiques propres de chacune.

Pour l'estimateur de HUBER, on prendra $c = k - .5\sqrt{k}$. Quand $k = 1$, l'estimateur se réduit au « Proposal 2 » de HUBER avec paramètre $\sqrt{.5} \approx .7$ ([5], p. 144). Pour des données normales, quand $k = 1$, en moyenne 48 % des individus auront un poids inférieur à 1 ; pour de grandes valeurs de k, ce pourcentage monte à 64 %. Pour l'estimateur biweight, on prendra $c = 2k + 2\sqrt{k}$. Quand $k = 1$, $c = 2^2$ et cet estimateur accorde un poids de 0 à toutes les observations à plus de deux écarts-types de la moyenne. Il devrait donc, en présence d'observations aberrantes, refléter la structure « principale » des données.

4.1. Un exemple simulé

Ici $k = 5$ et $n = 50$. Les 40 premières observations proviennent d'une loi normale multivariée de moyenne 0 et ayant une matrice de variances covariances dont les valeurs propres sont 1, 1, 1, 4 et 25 et les vecteurs propres correspondants :

$$\frac{1}{\sqrt{5}}\begin{bmatrix}1\\1\\1\\1\\1\end{bmatrix}, \frac{1}{\sqrt{2}}\begin{bmatrix}1\\-1\\0\\0\\0\end{bmatrix}, \frac{1}{\sqrt{6}}\begin{bmatrix}1\\1\\-2\\0\\0\end{bmatrix}, \frac{1}{\sqrt{12}}\begin{bmatrix}1\\1\\1\\-3\\0\end{bmatrix}, \frac{1}{\sqrt{20}}\begin{bmatrix}1\\1\\1\\1\\-4\end{bmatrix}$$

Les dix derniers points sont normaux, de matrice de variances covariances I.

Les moyennes des cinq premiers et des cinq derniers sont respectivement :

$$5\sqrt{2}\left[\frac{1}{\sqrt{5}}\begin{bmatrix}1\\1\\1\\1\\1\end{bmatrix}+\frac{1}{\sqrt{2}}\begin{bmatrix}1\\-1\\0\\0\\0\end{bmatrix}\right] \text{ et } 5\sqrt{2}\left[\frac{1}{\sqrt{5}}\begin{bmatrix}1\\1\\1\\1\\1\end{bmatrix}-\frac{1}{\sqrt{2}}\begin{bmatrix}1\\-1\\0\\0\\0\end{bmatrix}\right]$$

Ainsi les 10 dernières données « perturbent » le sous-échantillon principal.

Les résultats qui sont présentés dans le tableau I et sur la figure 1 sont clairs. L'analyse classique met en évidence les trois composantes de l'échantillon tandis que les procédures robustes reflètent la structure du sous-échantillon principal.

TABLEAU I

Matrices de corrélations et leurs trois premières valeurs propres pour l'exemple 1

analyse classique n = 50					analyse de HUBER n = 50				
1	−.41	.53	.29	.06	1	.35	.65	−.12	−.47
	1	.44	.32	.17		1	.57	−.11	−.54
		1	.56	.07			1	−.07	−.07
			1	.30				1	−.15
				1					1
valeurs propres			pourcentages cumulatifs		valeurs propres			pourcentages cumulatifs	
2.09			42		2.65			53	
1.43			70		1.08			75	
.95			89		.66			88	
analyse biweight n = 50					analyse classique du sous-échantillon principal n = 40				
1	.79	.78	−.35	−.69	1	.68	.63	−.30	−.59
	1	.65	−.17	−.70		1	.68	−.27	−.67
		1	−.18	−.86			1	−.19	−.76
			1	−.18				1	−.21
				1					1
valeurs propres			pourcentages cumulatifs		valeurs propres			pourcentages cumulatifs	
3.27			65		3.04			61	
1.14			88		1.17			84	
.38			96		.37			92	

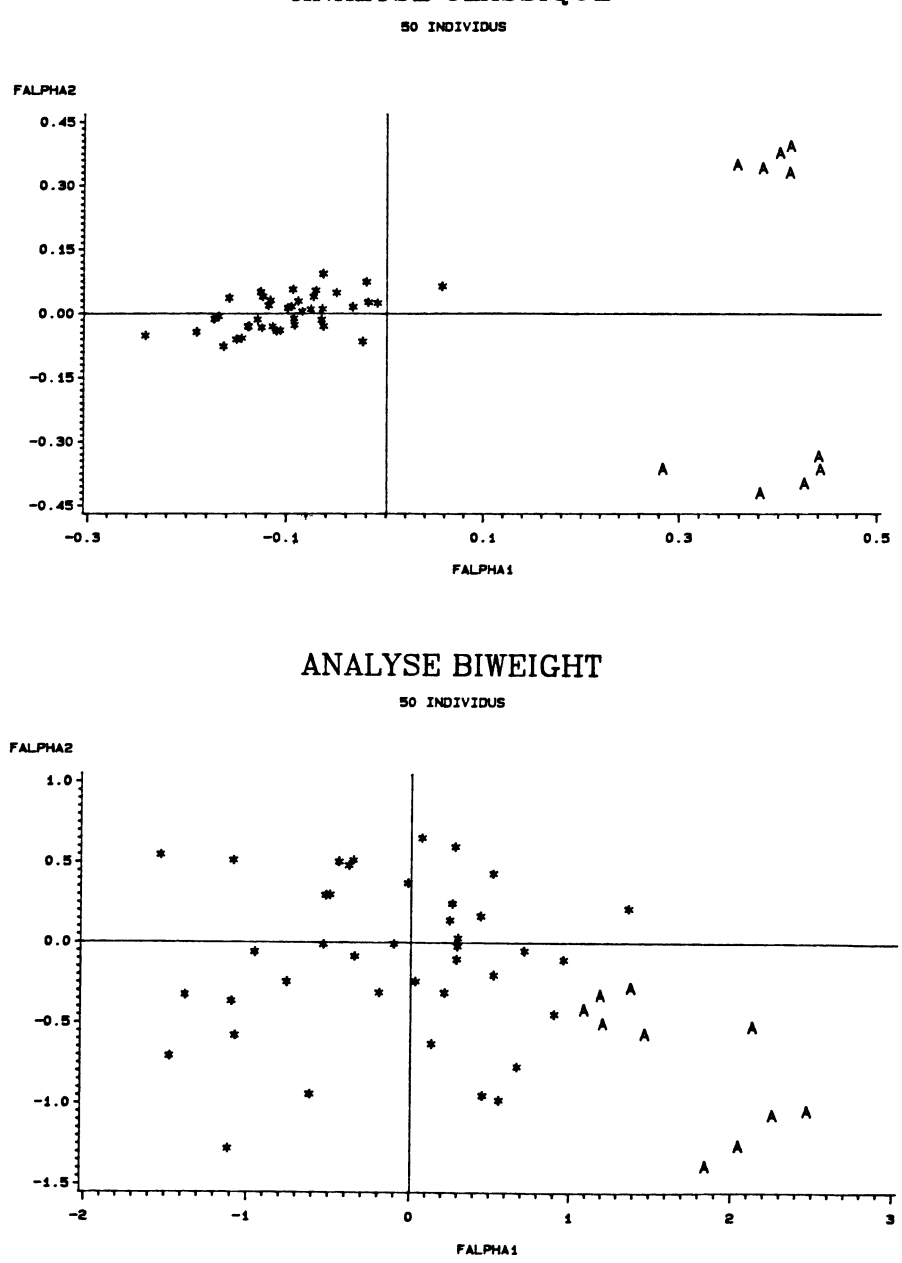

FIGURE 1

Analyse classique et biweight de l'exemple 1. Les individus du sous-échantillon principal sont représentés par une étoile; les dix autres par un A.

L'analyse biweight donne le poids 0 à 12 données, soit aux 10 dernières et aux 2 données les plus excentriques du sous-échantillon principal. Notons que pour expliquer 85 % de l'inertie elle requière deux axes tandis que l'analyse classique en prend 3. Ainsi en éliminant les individus éloignés du centre de gravité, l'analyse robuste révèle une structure plus simple. Evidemment l'étude de cette structure doit être complétée par un examen attentif des données éliminées.

4.2. Exemple de JAMBU ([11], p. 120)

Le nombre d'heures consacrés à dix activités différentes par 28 classes de sujets est analysé. Ces activités sont PROF, les activités professionnelles, TRAN, le transport, MENA, les travaux ménagers, ENFA, le temps consacré aux enfants, COUR, les courses, REPA, les repas, SOMM, le sommeil, TELE, la télévision et LOIS, les loisirs. Chaque classe de sujets est caractérisée par le pays d'origine et une catégorie socio-économique (voir le tableau II).

Le faible nombre d'individus (28) par rapport au nombre de variables (10) rend l'interprétation des poids hasardeuse : un individu peut recevoir un poids faible dans l'analyse robuste, en contribuant fortement dans l'analyse classique à un axe n'ayant aucune signification statistique. Par exemple, l'estimateur de HUBER donne un poids de .08 à FMWE (femme mariée de l'Europe de l'Ouest). Ce faible poids est, en grande partie, dû au fait que, dans l'analyse classique, la contribution de FMWE au 10^e axe est .51. Malgré tout, l'examen des poids de l'analyse biweight nous révèle que les catégories des femmes non actives et mariées reçoivent des poids pratiquement nuls dans chacun des quatre pays. De plus, la somme des sept poids accordés aux américains, .92, comparée à 2.16, 2.54 et 1.83 pour les trois autres pays, indique une différence de comportement entre américains et européens.

TABLEAU II

Caractères représentant, à la figure 2, les 28 classes de sujet.
Chaque classe est caractérisée par un pays d'origine et
une catégorie socio-économique

Pays d'origine	Hommes actifs (HA)	Femmes actives (FA)	Femmes non actives (FNA)	Hommes mariés (HM)	Femmes mariées (FM)	Hommes céli. (HC)	Femmes céli. (FC)
US	HAUS	FAUS	FNAU	HMUS	FMUS	HCUS	FCUS
Europe de l'ouest (WE)	HAWE	FAWE	FNAW	HMWE	FMWE	HCWE	FCWE
Yougoslavie (YO)	HAYO	FAYO	FNAY	HMYO	FMYO	HCYO	FCYO
Europe de l'est (ES)	HAES	FAES	FNAE	HMES	FMES	HCES	FCES

TABLEAU III

Matrices de corrélations classiques (première ligne) et biweight (deuxième ligne)

	Prof.	Men.	Enf.	Tél.	Loisir	Trans.	Course	Toilet.	Repas	Somm.
Act. prof.	1	−.91 −.79	−.86 −.52	−.05 −.39	−.25 .04	−.94 .60	−.65 −.59	−.11 −.57	−.46 .26	−.55 −.01
Travaux ménagers		1	.86 .90	−.21 −.56	−.05 −.46	−.87 −.62	.50 .33	−.04 .42	.36 −.28	.44 −.17
Enfant			1	.12 −.36	−.06 −.71	−.81 −.37	.54 .26	.12 .40	.36 −.31	.28 −.38
Télévision				1	−.07 −.28	−.04 .18	.21 −.06	.32 −.21	.32 .51	.02 .20
Loisir					1	−.16 .32	.24 .06	.06 −.02	.06 −.23	.27 .08
Transport						1	−.5 .14	−.08 .04	−.6 −.47	−.7 −.6
Course							1	.59 .75	−.18 −.68	−.02 −.57
Toilette								1	−.3 −.61	−.21 −.42
Repas									1	.82 .83
Sommeil										1

En résumé, la matrice des corrélations biweight représente la structure des données dans cinq des sept catégories socio-économiques (que nous appellerons « professionnelles »). De plus, les Etats-Unis participent moins à cette structure que les autres pays.

Les matrices de corrélations classique et biweight sont présentées au tableau III. La taille de l'échantillon étant faible, les résultats de l'analyse de HUBER sont très semblables à ceux de l'analyse classique et ne seront pas commentés. Un changement est digne de mention : la corrélation enfant-loisir passe de −.06 à −.71 : les « professionnels » qui consacrent du temps aux enfants n'ont plus de temps pour les loisirs !

Contrairement au premier exemple, l'emploi de méthodes robustes ne simplifie pas la structure des données : tant avec la méthode biweight qu'avec la méthode classique, les quatre premiers axes sont importants. Le tableau IV compare les qualités de représentation des individus selon les deux méthodes dans le plan des 2 premiers axes. Avec l'approche classique, ce plan représente certains individus très bien, d'autres très mal. L'approche biweight donne un

TABLEAU IV

Histogramme « Stem and Leaf » des mesures de qualités de représentation dans les deux premiers axes obtenus avec l'analyse classique et l'analyse biweight

Biweight		Classique	
0.0		0.0	53
.1	6	.1	57
.2	88	.2	1
.3	20	.3	735
.4	40	.4	31
.5	005	.5	01
.6	3696507178	.6	26
.7	20	.7	5977
.8	9103	.8	73849
.9	98	.9	15300

compromis; la plupart des individus sont relativement bien représentés. Pour seulement sept d'entre eux, la qualité de la représentation est inférieure à 50 %, comparativement à 10 pour la méthode classique. De plus, les sept individus mal représentés se répartissent en deux groupes distincts : les cinq catégories professionnelles américaines et les hommes actifs et mariés des pays de l'est.

L'interprétation du graphique des deux premiers axes biweight se fait en travaillant avec les droites $y = x$ et $y = -x$. La première oppose transport-activité professionnelle à travaux ménagers-enfants. Elle ordonne trois groupes de catégories socio-économiques selon le temps consacré aux activités professionnelles : les femmes mariées et non actives, les femmes actives et célibataires et les hommes. La deuxième oppose les binômes course-toilette et repas-sommeil. Elle sépare les pays en trois groupes, selon l'importance donnée à un binôme ou à l'autre : l'Europe de l'ouest, l'Europe de l'est et la Yougoslavie et finalement les américains, sauf chez les hommes où les deux derniers groupes se chevauchent. Un examen attentif des axes trois et quatre révèle que les cinq catégories américaines professionnelles sont mal représentées surtout à cause de leur grand usage de la télévision, ce qui n'apparait pas sur la figure 2. Le graphique correspondant pour l'A.C.P. classique se trouve à la page 123 de [11].

En conclusion, l'analyse biweight donne des résultats semblables à ceux de l'analyse classique. Cependant, elle fait ressortir plus clairement la séparation des individus en différents groupes sur les deux premiers axes.

5. Conclusion

L'analyse robuste est un complément de l'analyse en composantes principales classique. Son utilisation en présence de données hétérogènes peut mettre en lumière des caractéristiques que l'analyse classique avait passées sous silence. On peut aussi utiliser l'analyse robuste pour valider les résultats d'une A.C.P.

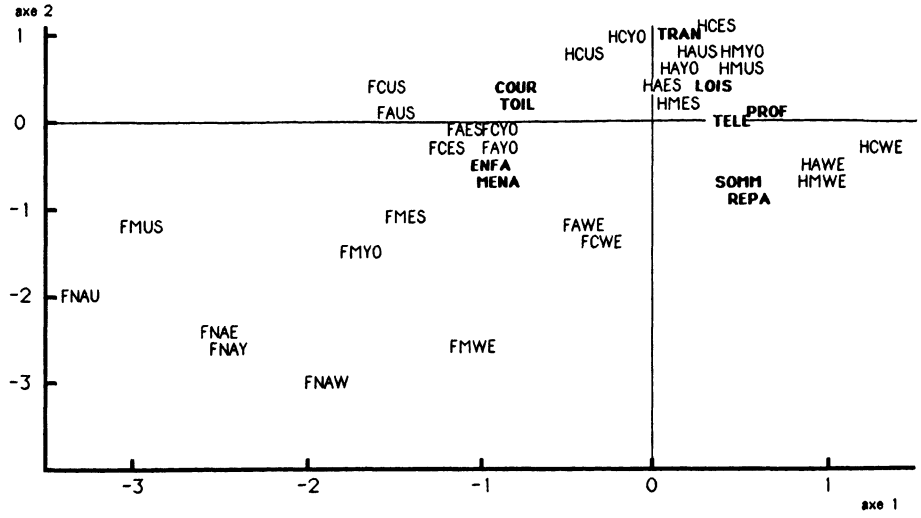

FIGURE 2

Représentation conjointe des classes d'individus (caractères ordinaires) et des activités (caractères gras) obtenue à l'aide d'une analyse en composantes principales robuste effectuée selon la méthode biweight.

ordinaire. Dans les cas où les deux méthodes donnent des résultats identiques, la crédibilité des conclusions s'en trouve augmentée.

Bibliographie

[1] J. BENASSENI. — Influence des poids des unités statistiques sur les valeurs propres en analyse en composantes principales. *Revue de Statistique Appliquée*, 1985, Vol. 33, p. 41-54.

[2] F. CRITCHLEY. — Influence in principal component analysis. *Biometrika*, 1985, Vol. 72, p. 627-636.

[3] D.A. FREEDMAN et P. DIACONIS. — On inconsistent M-estimators. *The Annals of Statistics*, Vol. 10, p. 454-461.

[4] F. HAMPEL, E. RONCHETTI, P.J. ROUSSEEUW et W. STAHEL. — *Robust Statistics. The Approach Based on Influence Functions.* John Wiley, 1986.

[5] P. HUBER. — *Robust Statistics.* John Willey, 1981.

[6] L. LEBART, A. MORINEAU et N. TABARD. — *Techniques de la Description Statistique*, Dunod 1977.

[7] R.A. MARONNA. — Robust M-estimators of multivariate location and scatter, *The Annals of Statistics*, 1976, Vol. 4, p. 51-67.

[8] S.M. STIGLER. — Simon Newcomb, Percy Daniell, and the history of robust estimation. *Journal of the American Statistical Association*, 1973, Vol. 68, p. 412-427.

[9] D. TYLER. — Robustness and efficiency properties of scatter matrices. *Biometrika*, 1983, Vol. 70, p. 411-420.

[10] D. TYLER. — Existence and uniqueness of the M-estimators of multivariate location and scatter, 1986, Prepublication, Département de statistiques, Rutgers University.

[11] N. VOLLE. — *Analyse des Données, 2^e édition*, Economica, 1981.