REVUE DE STATISTIQUE APPLIQUÉE

AZIZ LAZRAQ
ROBERT CLÉROUX

Analyse de la redondance robuste

Revue de statistique appliquée, tome 53, n° 2 (2005), p. 43-65
<http://www.numdam.org/item?id=RSA_2005__53_2_43_0>

Rev. Statistique Appliquée, 2005, **LIII** (2), 43-65

ANALYSE DE LA REDONDANCE ROBUSTE

Aziz LAZRAQ[(1)], Robert CLÉROUX[(2)]

(1) École Nationale de l'Industrie Minérale, Rabat
courriel : lazraq@enim.ac.ma
(2) Département de mathématiques et de statistique, Université de Montréal
courriel : cleroux@dms.umontreal.ca

RÉSUMÉ

Dans cet article on s'intéresse à l'analyse de la redondance robuste. On estime la matrice de covariance par un estimateur RMCD et on obtient les paramètres de l'analyse de la redondance ainsi que leurs fonctions d'influence. On compare ces dernières à celles obtenues dans le cas classique c'est-à-dire lorsque l'estimateur habituel de la matrice de covariance est utilisé. On refait la même démarche dans la situation où l'analyse de la redondance est effectuée à partir d'une matrice de corrélation. Un exemple compare les démarches robuste et classique et une étude de simulation montre clairement l'avantage de l'approche robuste en présence de données contaminées.

Mots-clés : *Analyse de la redondance, données contaminées, estimateurs robustes.*

ABSTRACT

This paper is concerned with robust redundancy analysis. The covariance matrix is estimated with an RMCD estimator and the parameters of redundancy analysis are obtained together with their influence functions. These functions are then compared with those obtained in the classical situation which uses the usual covariance matrix estimator. The same process is repeated for the case where redundancy analysis is made from the correlation matrix. An example compares the robust and the classical approaches and a simulation study shows clearly the advantage of the robust approach in the presence of contaminated data.

Keywords : *Redundancy analysis, contaminated data, robust estimators.*

1. Introduction

Lorsque l'on cherche à prédire un ensemble Y de variables à partir d'un ensemble X de prédicteurs on utilise habituellement l'indice de redondance de Stewart et Love (1968) généralisé par la suite par Gleason (1976). Cet indice est la fraction de la variance totale de Y expliquée par X linéairement et mesure donc la qualité de la prédiction de Y par X. Utilisant cette mesure de redondance, van den Wollenberg (1977) a introduit l'analyse de la redondance qui consiste à extraire des facteurs ou des variables latentes à partir de l'ensemble X qui maximisent la redondance.

C'est en quelque sorte un compromis entre l'analyse canonique et la régression linéaire multivariée. L'analyse de la redondance est aussi en relation avec l'analyse en composantes principales par rapport à un ensemble de variables instrumentales (voir Rao, 1964 ou Bry, 1996). Lazraq et Cléroux (2002A) présentent une méthode inférentielle de sélection des variables latentes sous l'hypothèse que $\begin{pmatrix} Y \\ X \end{pmatrix}$ possède une loi normale multivariée. On y trouve aussi une bibliographie des auteurs ayant traité de l'analyse de la redondance.

Dans cet article on s'intéresse à l'analyse de la redondance robuste. On travaille à partir d'un estimateur robuste de la matrice de covariance (estimateur RMCD, Reweighted Minimum Covariance Determinant) pour obtenir les paramètres de l'analyse de la redondance. On obtient par la suite les fonctions d'influence de ces paramètres. On obtient également les fonctions d'influence lorsque l'analyse de la redondance est effectuée à partir de la matrice de corrélation robuste. Ces fonctions d'influence sont ensuite comparées à celles que l'on obtient dans le cas classique, c'est-à-dire à partir de la matrice de covariance ou de corrélation empirique classique. Un exemple contenant plusieurs données aberrantes compare l'approche robuste et l'approche classique. Finalement une étude de simulation indique dans quelles circonstances l'une est supérieure à l'autre pour l'estimation des paramètres.

Cet article s'inspire de celui de Croux et Dehon (2002) qui introduit l'analyse canonique robuste. Nous utilisons la même approche et autant que possible la même notation.

L'article est organisé comme suit. Nous rappelons l'analyse de la redondance à la Section 2. La Section 3 rappelle les estimateurs MCD et RMCD. Dans la Section 4 nous présentons l'analyse de la redondance basée sur une matrice de covariance robuste et obtenons les fonctions d'influence des paramètres dans la Section 5. La Section 6 est consacrée aux fonctions d'influence obtenues à partir d'une matrice de corrélation robuste. Suivent un exemple à la Section 7 et une étude de simulation à la Section 8. Finalement la Section 9 contient la conclusion de l'article. Les preuves des résultats de la Section 6 sont reportées en annexe.

2. Analyse de la redondance

On résume la méthode de l'analyse de la redondance telle que présentée dans Lazraq et Cléroux (2002A). Le problème est de prédire un ensemble de variables dites dépendantes $Y = (Y_1, Y_2, \ldots, Y_p)'$ à partir d'un ensemble de facteurs extraits de variables dites indépendantes $X = (X_1, X_2, \ldots, X_q)'$.

Le vecteur aléatoire $\begin{pmatrix} Y \\ X \end{pmatrix}$: $m \times 1$ où $m = p + q$ est supposé multinormal de moyenne $\mu = \begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix}$ et de matrice de covariance $\Sigma = \begin{pmatrix} \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_{XX} \end{pmatrix}$ où μ_Y : $p \times 1$, μ_X : $q \times 1$, Σ_{YY} : $p \times p$, Σ_{XX} : $q \times q$ et $\Sigma_{XY} = \Sigma'_{YX}$: $q \times p$.

On suppose que Σ_{XX} est inversible. La méthode utilise l'indice de redondance ρI défini par

$$\rho I(Y,X) = \frac{\mathrm{tr}(\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY})}{\mathrm{tr}(\Sigma_{YY})} \tag{1}$$

et qui peut aussi être écrit sour la forme

$$\rho I(Y,X) = \frac{\Sigma_{i=1}^{p}\sigma_{Y_i}^2 \bar{R}^2(Y_i;X_1,X_2,\ldots,X_q)}{\Sigma_{i=1}^{p}\sigma_{Y_i}^2} \tag{2}$$

où $\bar{R}^2(Y_j;X_1,X_2,\ldots,X_q)$ est le carré du coefficient de corrélation multiple entre Y_j, la j^e composante de Y, et les prédicteurs $X_1,X_2,\ldots,X_q$ et où $\sigma_{Y_j}^2$ est la variance de Y_j. Ainsi, l'indice de redondance est la fraction de la variance totale de Y qui peut être expliquée par la régression linéaire multivariée de Y sur X. Plus de détails sur les propriétés de $\rho I(Y,X)$ ou sur d'autres mesures de corrélation vectorielle ou de redondance peuvent être obtenues dans Cramer et Nicewander (1979) et Lazraq et Cléroux (1988).

Soit $s = \min(p,q)$. Dans l'analyse de la redondance on cherche $r(r \leqslant s)$ fonctions linéaires non corrélées $t_j = \alpha_j' X$, $j = 1,2,\ldots,r$, qui maximisent $\sum_{j=1}^{r}\rho I(Y,t_j)$ sujet à $\mathrm{var}(t_j) = \alpha_j'\Sigma_{XX}\alpha_j = 1$ pour chaque j. Ici $\rho I(Y,t_j)$ est la fraction de la variance totale de Y expliquée par la régression linéaire multivariée de Y sur t_j. À partir de (2) on peut écrire

$$\sum_{j=1}^{r}\rho I(Y,t_j) = \frac{\displaystyle\sum_{j=1}^{r}\sum_{i=1}^{p}\sigma_{Y_i}^2\mathrm{corr}^2(Y_i,\alpha_j'X)}{\displaystyle\sum_{i=1}^{p}\sigma_{Y_i}^2}. \tag{3}$$

Le dénominateur de (3) ne dépend pas de α_j et son numérateur est égal à

$$\begin{aligned}\sum_{j=1}^{r}\sum_{i=1}^{p}\sigma_{Y_i}^2\mathrm{corr}^2(Y_i,\alpha_j'X) &= \sum_{j=1}^{r}\sum_{i=1}^{p}\sigma_{Y_i}^2\frac{(\alpha_j'\Sigma_{XY_i}\Sigma_{Y_iX}\alpha_j)}{\sigma_{Y_i}^2\alpha_j'\Sigma_{XX}\alpha_j}\\ &= \sum_{j=1}^{r}\sum_{i=1}^{p}\alpha_j'\Sigma_{XY_i}\Sigma_{Y_iX}\alpha_j \text{ puisque } \alpha_j'\Sigma_{XX}\alpha_j = 1, \forall j\\ &= \sum_{j=1}^{r}\alpha_j'\left[\sum_{i=1}^{p}\Sigma_{XY_i}\Sigma_{Y_iX}\right]\alpha_j = \sum_{j=1}^{r}\alpha_j'\Sigma_{XY}\Sigma_{YX}\alpha_j. \end{aligned} \tag{4}$$

Donc le problème de l'analyse de la redondance se réduit à la maximisation de (4) sous la contrainte $\alpha_j'\Sigma_{XX}\alpha_j = 1$ pour chaque j. Si l'on pose maintenant $c_j = \Sigma_{XX}^{1/2}\alpha_j$, $j = 1, 2, \ldots, r$, alors (4) devient

$$\sum_{j=1}^{r} c_j' \left(\Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YX}\Sigma_{XX}^{-1/2} \right) c_j \tag{5}$$

et la condition $\alpha_j'\Sigma_{XX}\alpha_j = 1$ devient $c_j'c_j = 1$. Puisque la matrice $\Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YX}\Sigma_{XX}^{-1/2}$ est symétrique, ses vecteurs propres sont orthogonaux. Il est connu (voir par exemple Rao (1973)) que (5) est maximisé lorsque $c_1, c_2, \ldots, c_r$, sont les vecteurs propres orthonormés correspondant aux valeurs propres $\lambda_1, \lambda_2, \ldots, \lambda_r$ de la matrice $\Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YX}\Sigma_{XX}^{-1/2}$ et alors le maximum de (5) quand $r = s$ est donné par

$$\begin{aligned} \sum_{j=1}^{s} \lambda_j &= \mathrm{tr}(\Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YX}\Sigma_{XX}^{-1/2}) = \mathrm{tr}(\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YX}) \\ &= \mathrm{tr}(\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}) \end{aligned} \tag{6}$$

c'est-à-dire que le maximum de (3) quand $r = s$ est $\rho I(Y, X)$. On peut maintenant démontrer la proposition suivante.

PROPOSITION 1. – *Pour chaque $j = 1, 2, \ldots, s$, la valeur propre λ_j associée au vecteur propre α_j, est la j^e plus grande valeur propre de $\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YX}$.*

Démonstration. – Nous avons, pour chaque $j = 1, 2, \ldots, s$,

$$\Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YX}\Sigma_{XX}^{-1/2}c_j = \lambda_j c_j.$$

Remplaçons c_j par sa valeur $\Sigma_{XX}^{1/2}\alpha_j$, on obtient

$$\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YX}\alpha_j = \lambda_j\alpha_j. \qquad \square$$

Remarquons également que, comme attendu,

$$\mathrm{cov}(t_h, t_\ell) = \alpha_h'\Sigma_{XX}\alpha_\ell = c_h'c_\ell = \delta_{h\ell}$$

où $\delta_{h\ell}$ est le delta de Kronecker, démontrant l'indépendance, au niveau de la population, entre les facteurs successifs de l'analyse de redondance. Dans Lazraq et Cléroux (2002A) il est démontré aussi que

$$\rho I(Y, t_j) = \frac{\lambda_j}{\alpha_j'\Sigma_{XX}\alpha_j \mathrm{tr}(\Sigma_{YY})} = \frac{\lambda_j}{\mathrm{tr}(\Sigma_{YY})} \tag{7}$$

c'est-à-dire que la mesure de dépendance entre Y et t_j est proportionnelle à λ_j.

3. Rappel des estimateurs MCD et RMCD

Ces estimateurs furent introduits dans la littérature par Rousseeuw (1985). Considérons un échantillons de n observations $Z_1, Z_2, \ldots, Z_n$ du vecteur $Z = (Y', X')'$ et notons par $d(Z_i, t, V) = [(Z_i - t)'V^{-1}(Z_i - t)]^{1/2}$ la distance entre Z_i et t par rapport à la métrique induite par la matrice définie positive V. Les estimateurs RMCD (Reweighted Minimum Covariance Determinant) de position t_n et de la matrice de covariance sont définis respectivement par

$$t_n = \frac{\sum_{i=1}^{n} w_i Z_i}{\sum_{i=1}^{n} w_i}, \quad C_n = c_1 \frac{\sum_{i=1}^{n} w_i (Z_i - t_n)(Z_i - t_n)'}{\sum_{i=1}^{n} w_i} \tag{8}$$

avec

$$C_n = \begin{pmatrix} C_{YY}^{(n)} & C_{YX}^{(n)} \\ C_{XY}^{(n)} & C_{XX}^{(n)} \end{pmatrix}, \; C_{YY}^{(n)} : p \times p, \; C_{XX}^{(n)} : q \times q, \; C_{XY}^{(n)} = C_{YX}^{(n)'} : q \times p$$

et où $c_1 = (1 - \delta) F_{\chi^2_{m+2}}(q\delta)$ est un facteur de convergence pour une distribution normale, $F_{\chi^2_{m+2}}$ est la fonction de répartition de la loi du chi-deux avec $m+2$ degrés de liberté et $q_\delta = \chi^2_{m,1-\delta}$, le quantile $1-\delta$ de la loi du chi-deux à m degrés de liberté. Les poids w_i sont donnés par

$$w_i = \begin{cases} 1 & \text{si} \quad d^2(Z_i, t_n^o, C_n^o) \leqslant q_\delta \\ 0 & \text{sinon} \end{cases} \tag{9}$$

où (t_n^o, C_n^o) sont les estimateurs initiaux MCD (Minimum Covariance Determinant). Pour déterminer les estimateurs MCD on considère tous les sous-ensembles de taille $h \leqslant n$ de l'échantillon de départ et on conserve les sous-ensembles dont le déterminant de la matrice de covariance est le plus petit. Les estimateurs (t_n^o, C_n^o) sont alors la moyenne classique et la matrice de covariance classique (à une constante multiplicative près) calculées à partir du sous-ensemble de h points optimaux au sens du déterminant minimal. En général la taille du sous-ensemble est égal à $h = \lfloor n(1-\alpha) \rfloor$ où $\alpha = 0.50$ ou 0.25 représente le point de rupture des estimateurs MCD ou RMCD. Pour plus de détails sur les estimateurs MCD et RMCD, leurs propriétés et des algorithmes de calcul, nous référons le lecteur aux article de Croux et Haesbroeck (2000) et Croux et Dehon (2002) ainsi qu'aux références qu'ils contiennent.

4. Analyse de la redondance basée sur une matrice de covariance robuste

On suppose que l'échantillon aléatoire $Z_1, Z_2, \ldots, Z_n$ provient de la loi $F = N(\mu, \Sigma)$ où $\mu : m \times 1$ et où $\Sigma \in MSPD$, l'ensemble des matrices $m \times m$

symétriques et définies positives. Soit $\mathcal{F}$ l'ensemble de toutes les lois sur $\mathbb{R}^m$ ou un large sous-ensemble de celui-ci. Croux et Dehon (2002) et aussi Croux et Haesbroeck (2000) introduisent les fonctionnelles statistiques correspondant à un estimateur de dispersion comme une application $C : \mathcal{F} \to MSPD$ qui envoie $G \in \mathcal{F}$ sur $C(G)$. Nous cherchons donc à estimer $C(F)$ et nous utiliserons librement les expressions «fonctionnelle C» ou «estimateurs C». Nous pourrons utiliser aussi $C(Z)$ à la place de $C(G)$ si Z suit la loi G. Notons qu'un estimateur naturel $C(F)$ est $C_n = C(F_n)$ où F_n est la fonction de répartition empirique.

La recherche des coefficients α_j (loadings) de la composante t_j de l'analyse de la redondance ainsi que la mesure λ_j associée consiste à trouver les vecteurs propres α_j et les valeurs propres λ_j de $\mathcal{A} = \Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YX}$. Les fonctionnelles associées aux estimateurs des valeurs propres seront notées ℓ_j et celles associées aux estimateurs des vecteurs propres a_j, $j = 1, 2, \ldots, q$. Il suit que pour $G \in \mathcal{F}$, $\ell_j(G)$ et $a_j(G)$ sont les valeurs et vecteurs propres de $A(G) = C_{XX}^{-1}(G)C_{XY}(G)C_{YX}(G)$ avec $C(G) = \begin{pmatrix} C_{YY}(G) & C_{YX}(G) \\ C_{XY}(G) & C_{XX}(G) \end{pmatrix}$ et $C_{YY}(G) : p \times p$, $C_{XX}(G) : q \times q$, $C_{XY}(G) = C'_{YX}(G) : q \times p$.

On suppose que l'estimateur C converge au sens de Fisher pour Σ et F, c'est-à-dire que $C(F) = \Sigma$, et qu'il est affine équivariant, c'est-à-dire que $C(AZ + b) = AC(Z)A'$ pour toute matrice $A : m \times m$ non singulière et tout vecteur $b : m \times 1$. Ces hypothèses impliquent la convergence de a_j et ℓ_j au sens de Fisher : $a_j(F) = \alpha_j$ et $\ell_j(F) = \lambda_j$. Aussi $C_{XX}(F) = \Sigma_{XX}$, $C_{YY}(F) = \Sigma_{YY}$ et $C_{YX}(F) = \Sigma_{YX}$.

En outre, si à partir de l'estimateur C de la matrice de dispersion on définit

$$\rho IC(G) = \rho IC(Y, X) = \frac{\mathrm{tr}(C_{YX}C_{XX}^{-1}C_{XY})}{\mathrm{tr}(C_{YY})}, \tag{10}$$

la valeur de ρI calculée à partir de la matrice $C(G)$, alors l'estimateur de ρIC en C_n sera noté

$$RI(C_n) = \rho IC(F_n) = \frac{\mathrm{tr}(C_{YX}^{(n)}C_{XX}^{(n)-1}C_{XY}^{(n)})}{\mathrm{tr}(C_{YY}^{(n)})}. \tag{11}$$

Enfin rappelons la définition de la fonction d'influence telle qu'introduite par Hampel (1974) et qui sera fort utile par la suite. Soit F une fonction de répartition et $F_\epsilon = (1 - \epsilon)F + \epsilon\delta_x$ une perturbation de F par δ_x, la fonction de répartition qui affecte la probabilité 1 en x. Soit T une fonctionnelle statistique. Alors la fonction d'influence de T au point Z en la distribution F et définie par

$$I(Z, T, F) = \lim_{\epsilon \to 0} \frac{T(F_\epsilon) - T(F)}{\epsilon} = \frac{\delta}{\delta\epsilon} T(F_\epsilon) \mid_{\epsilon=0} . \tag{12}$$

La fonction d'influence de $RI(C_n)$ ainsi que ses distributions sous H_0 : pas de relation linéaire entre Y et X, et sous H_1 : non H_0, sont données dans Lazraq et Cléroux (2002 B) pour le cas des S-estimateurs et pour les lois F elliptiques. Des résultats analogues peuvent facilement être obtenus pour le cas des estimateurs RMCD. Nous

nous intéresserons, dans la section suivante, aux fonctions d'influence des estimateurs a_j et ℓ_j. Nous aurons besoin des deux lemmes suivants que l'on trouve dans Croux et Dehon (2002).

LEMME 1. – *Pour toute fonctionnelle C associée à un estimateur multivarié de dispersion affine équivariant possédant une fonction d'influence, il existe deux fonctions γ_C et $\delta_C : [0, \infty) \to \mathbb{R}$ telles que*

$$IF(Z, C, F) = \gamma_C(d(Z))(Z - \mu)(Z - \mu)' - \delta_C(d(Z))\Sigma \tag{13}$$

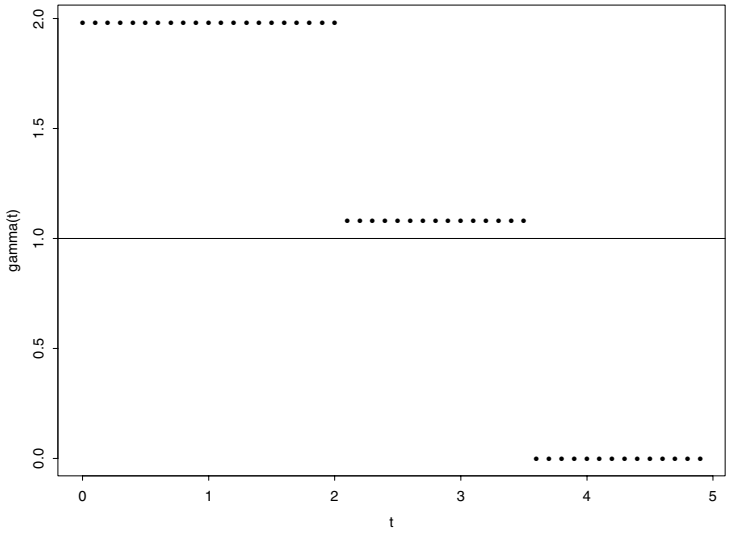
où $d^2(Z) = (Z - \mu)'\Sigma^{-1}(Z - \mu)$ et $F = N(\mu, \Sigma)$.

LEMME 2. – *Soient (λ_j, α_j) les paires de valeurs et vecteurs propres d'une matrice $\mathcal{A}$ quelconque telles que $\alpha_i'\Sigma\alpha_j = \delta_{ij}$ pour $i, j = 1, 2. \ldots, p$ et Σ une matrice définie positive quelconque. Alors*

$$IF(Z, \ell_j, F) = \alpha_j'\Sigma IF(Z, A, F)\alpha_j \tag{14}$$

$$IF(Z, a_j, F) = \sum_{\substack{k=1 \\ k \neq j}}^{p} \left(\frac{\alpha_k'\Sigma IF(Z, A, F)\alpha_j}{\lambda_j - \lambda_k} \right) \alpha_k - \frac{1}{2}(\alpha_j' IF(Z, C, F)\alpha_j)\alpha_j \tag{15}$$

où A est une fonctionnelle telle que $A(F) = \mathcal{A}$, ℓ_j et a_j sont les valeurs et vecteurs propres de A avec $a_j'Ca_j = 1$ et C est une fonctionnelle telle que $C(F) = \Sigma$.



GRAPHIQUE 1
Fonctions $\gamma_{\mathrm{COV}}(F)$ en trait continu et $\gamma_{\mathrm{RMCD}}(F)$ en trait pointillé pour $m = 5$

Les fonctions γ_C et δ_C sont données dans Croux et Haesbroeck(2000). Pour l'estimateur RMCD, la fonction γ_C joue un rôle plus important que δ_C. C'est une fonction en escalier possédant deux discontinuités. Les deux fonctions sont non croissantes et deviennent nulles après un certain point. Leur contribution à la fonction d'influence décroît si la distance entre Z et μ par rapport à la métrique induite par Σ augmente. Par contre, la fonction γ_{COV}, où COV est l'estimateur classique de la matrice de covariance, est constante et égale à 1 : elle donne le même poids à tous les points. Les propriétés précédentes sont tirées de Croux et Dehon (2002) et sont confirmées par le Graphique 1 qui montre les fonctions $\gamma_{\mathrm{COV}}(F)$ en trait continu et $\gamma_{\mathrm{RMCD}}(F)$ avec $m = 5$ en trait pointillé.

5. Fonctions d'influence des paramètre ℓ_j et a_j de l'analyse de la redondance

Le lemme et les deux théorèmes suivants nous fourniront les fonctions d'influence cherchées.

LEMME 3. – *La fonction d'influence de $A = C_{XX}^{-1} C_{XY} C_{YX}$ au point Z, en la distribution F est donnée par*

$$\begin{aligned} IF(Z, A, F) = \gamma_C(d(Z))\Sigma_{XX}^{-1}\big[&(X-\mu_X)(Y-\mu_Y)'\Sigma_{YX} \\ &+ \Sigma_{XY}(Y-\mu_Y)(X-\mu_X)' \\ &- (X-\mu_X)(X-\mu_X)'\mathcal{A}\big] - \delta_C(d(Z))\mathcal{A} \end{aligned} \tag{16}$$

où $\mathcal{A} = \Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YX}$.

Démonstration. – Par définition

$$\begin{aligned} IF(Z, A, F) &= \frac{\partial}{\partial \varepsilon} A(F_\varepsilon) \mid_{\varepsilon=0} \\ &= IF(Z, C_{XX}^{-1}, F)\Sigma_{XY}\Sigma_{YX} + \Sigma_{XX}^{-1} IF(Z, C_{XY}, F)\Sigma_{YX} \\ &\quad + \Sigma_{XX}^{-1}\Sigma_{XY} IF(Z, C_{YX}, F). \end{aligned} \tag{17}$$

Mais ces différentes fonctions d'influence s'écrivent (voir Lazraq et Cléroux, 2002B)

$$\begin{aligned} IF(Z, C_{XX}^{-1}, F) &= -\Sigma_{XX}^{-1} IF(Z, C_{XX}, F)\Sigma_{XX}^{-1} \\ &= -\Sigma_{XX}^{-1}[\gamma_C(d(Z))(X-\mu_X)(X-\mu_X)' \\ &\quad - \delta_C(d(Z))\Sigma_{XX}]\Sigma_{XX}^{-1} \end{aligned} \tag{18}$$

$$IF(Z, C_{XY}, F) = \gamma_C(d(Z))(X-\mu_X)(Y-\mu_Y)' - \delta_C(d(Z))\Sigma_{XY} \tag{19}$$

$$IF(Z, C_{YX}, F) = IF'(Z, C_{XY}, F). \tag{20}$$

En reportant (18), (19) et (20) dans (17) on obtient le résultat désiré. □

THÉORÈME 1. – *La fonction d'influence de ℓ_j basé sur un estimateur C affin équivariant de la matrice de covariance en la loi $F = N(\mu, \Sigma)$ est*

$$IF(Z, \ell_j, F) = \gamma_C(d(Z))\big[2B_j'(Y - \mu_Y)t_j - \lambda_j t_j^2\big] - \delta_C(d(Z))\lambda_j \tag{21}$$

où $B_j = \Sigma_{YX}\alpha_j$ est le vecteur des coefficients de régression linéaire de Y sur t_j, $t_j = \alpha_j'(X - \mu_X)$, $d^2(Z) = (Z - \mu)'\Sigma^{-1}(Z - \mu)$ et $\Sigma = C(F)$.

Démonstration. – En notant que $\alpha_i'\Sigma_{XX}\alpha_j = \delta_{ij}$ et en utilisant le Lemme 2 on obtient

$$IF(Z, \ell_j, F) = \alpha_j'\Sigma_{XX} IF(Z, A, F)\alpha_j \tag{22}$$

et par le Lemme 3

$$\begin{aligned} IF(Z, \ell_j, F) = \gamma_C(d(Z))\big[&\alpha_j'(X - \mu_X)(Y - \mu_Y)'\Sigma_{YX}\alpha_j \\ &+ \alpha_j'\Sigma_{XY}(Y - \mu_Y)(X - \mu_X)'\alpha_j \\ &- \alpha_j'(X - \mu_X)(X - \mu_X)'\mathcal{A}\alpha_j\big] - \delta_C(d(Z))\alpha_j'\Sigma_{XX}\mathcal{A}\alpha_j. \end{aligned} \tag{23}$$

En notant que $\mathcal{A}\alpha_j = \lambda_j\alpha_j$ et en simplifiant, (23) devient

$$IF(Z, \ell_j, F) = \gamma_C(d(Z))\big[2\alpha_j'\Sigma_{XY}(Y - \mu_Y)t_j - \lambda_j t_j^2\big] - \lambda_j\delta_C(d(X)). \tag{24}$$

Or le coefficient de régression linéaire de Y sur t_j est donné par $B_j = \Sigma_{Yt_j}\mathrm{var}^{-1}(t_j)$ et comme $\Sigma_{Yt_j} = \mathrm{cov}(Y, \alpha_j'(X - \mu_X)) = \Sigma_{YX}\alpha_j$ et $\mathrm{var}(t_j) = \alpha_j'\Sigma_{XX}\alpha_j = 1$, on obtient le résultat. □

THÉORÈME 2. – *Dans le même contexte, la fonction d'influence de a_j est donnée par*

$$\begin{aligned} IF(Z, a_j, F) = \sum_{\substack{k=1 \\ k\neq j}}^{s} \frac{1}{\lambda_j - \lambda_k}\Big\{&\gamma_C(d(Z))[Y - \mu_Y)'B_j t_k + B_k'(Y - \mu_Y)t_j \\ &- \lambda_j t_k t_j]\Big\}\alpha_k - \frac{1}{2}\big[\gamma_C(d(Z))t_j^2 - \delta_C(d(Z)\big]\alpha_j \end{aligned} \tag{25}$$

où $s = \min(p, q)$

Démonstration. – Par le Lemme 2 on peut écrire

$$IF(Z,a_j,F)=\sum_{\substack{k=1\\k\neq j}}^{s}\Big(\frac{\alpha_k'\Sigma_{XX}IF(Z,A,F)\alpha_j}{\lambda_j-\lambda_k}\Big)\alpha_k-\frac{1}{2}(\alpha_j'IF(Z,C_{XX},F)\alpha_j)\alpha_j. \tag{26}$$

Or nous avons, par le Lemme 3, pour $k\neq j$, en utilisant $t_j=\alpha_j'(X-\mu_X)$ et compte tenu de ce que $\alpha_j'\Sigma_{XX}\alpha_k=0$

$$\begin{aligned}\alpha_k'\Sigma_{XX}IF(Z,A,F)\alpha_j&=\gamma_C(d(Z))\\&\quad\big[t_k(Y-\mu_Y)'\Sigma_{YX}\alpha_j+\alpha_k'\Sigma_{XY}(Y-\mu_Y)t_j-\lambda_jt_kt_j\big]\\&=\gamma_C(d(Z))\big[t_k(Y-\mu_Y)'B_j+B_k'(Y-\mu_Y)t_j-\lambda_jt_kt_j\big].\end{aligned} \tag{27}$$

D'autre part nous avons aussi pour le second terme de (26), en utilisant le Lemme 1,

$$\begin{aligned}\alpha_j'IF(Z,C_{XX},F)\alpha_j&=\gamma_C(d(Z))\big[\alpha_j'(X-\mu_X)(X-\mu_X)'\alpha_j\big]\\&\quad-\delta_C(d(Z))\alpha_j'\Sigma_{XX}\alpha_j\\&=\gamma_C(d(Z))t_j^2-\delta_C(d(Z))\end{aligned} \tag{28}$$

car $\alpha_j'\Sigma_{XX}\alpha_j=1$.

Pour compléter la preuve on reporte (27) et (28) dans (26). □

Remarquons que les fonctions d'influence ℓ_j^{cov} et a_j^{cov} basées sur l'estimateur classique de la matrice de covariance s'obtiennent en posant $\gamma_C\equiv 1$ et $\delta_C\equiv 1$ dans les deux formules précédentes (21) et (25). Par exemple, à partir du théorème 1, on a

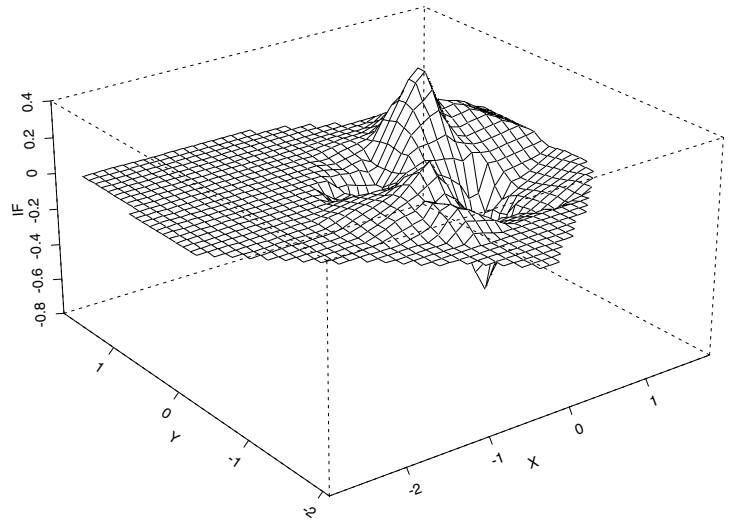
$$IF(Z,\ell_j^{\mathrm{cov}},F)=2B_j'(Y-\mu_Y)t_j-\lambda_jt_j^2-\lambda_j$$

ce qui permet d'écrire

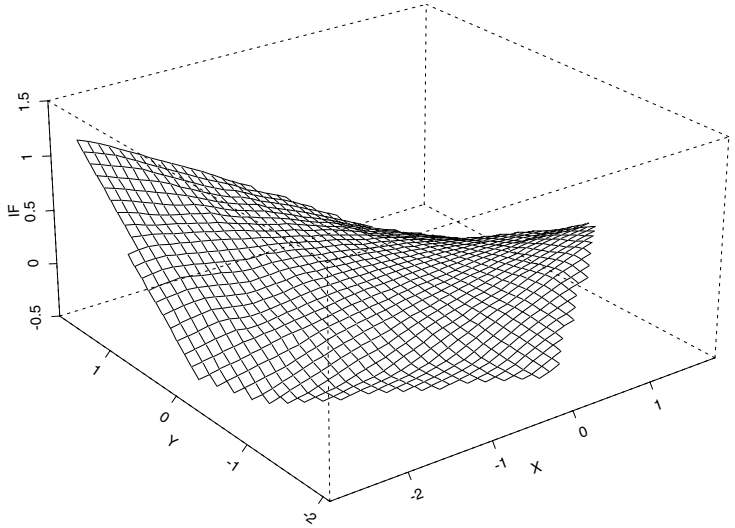
$$IF(Z,\ell_j^{\mathrm{RMCD}},F)=\gamma_C(d(Z))IF(Z,\ell_j^{\mathrm{cov}},F)+[\gamma_C(d(Z))-\delta_C(d(Z))]\lambda_j.$$

On voit bien l'importance de la fonction $\gamma_C(d(Z))$ qui est en quelque sorte une fonction de sous-pondération qui permet de garder $IF(Z,\ell_j,F)$ bornée. Un commentaire analogue suit la relation entre $IF(Z,a_j,F)$ et $IF(Z,a_j^{\mathrm{cov}},F)$.

Les graphiques 2 et 3 montrent respectivement dans des régions finies les fonctions d'influence $IF(Z,\ell_1^{\mathrm{RMCD}},F)$ et $IF(Z,\ell_1^{\mathrm{cov}},F)$ obtenues à partir d'un échantillon aléatoire de taille $n=100$ de $Z\sim N_2(0,\Sigma)$ avec $\Sigma=\begin{pmatrix}0,7 & 0\\ 0 & 1\end{pmatrix}$, $p=q=1$. On remarque en particulier que $IF(Z,\ell_1^{\mathrm{RMCD}},F)$ devient constante et bornée. On ne peut cependant en dire autant de $IF(Z,\ell_1^{\mathrm{cov}},F)$. Cela indique qu'un point aberrant pourrait avoir une influence importante sur l'estimateur.



GRAPHIQUE 2
Fonction d'influence $IF(Z, \ell_1^{\mathrm{RMCD}}, F)$



GRAPHIQUE 3
Fonction d'influence $IF(Z, \ell_1^{\mathrm{cov}}, F)$

6. Fonctions d'influence basées sur la matrice de corrélation

Puisque les résultats d'une analyse de la redondance dépendent des échelles de mesure des variable originelles, il est fréquent d'utiliser la matrice de corrélation

plutôt que la matrice de covariance. L'analyse de la redondance robuste basée sur la matrice de corrélation est donc pertinente.

Nous utiliserons la notation suivante :

i) pour toute matrice $E : m \times m$ on notera $\mathrm{diag}(E)$ la matrice diagonale $m \times m$ constituée uniquement des éléments diagonaux de E

ii) la matrice de corrélation de la population mère s'écrit $P = \Sigma_D^{-1/2}\Sigma\Sigma_D^{-1/2}$ où $\Sigma_D = \mathrm{diag}(\Sigma)$ et on écrira $\Sigma_{DX} = \mathrm{diag}(\Sigma_{XX})$, $\Sigma_{DY} = \mathrm{diag}(\Sigma_{YY})$ et $P = \begin{pmatrix} P_{YY} & P_{YX} \\ P_{XY} & P_{XX} \end{pmatrix}$ avec $P_{YY} : p \times p, P_{XX} : q \times q, P_{YX} = P'_{XY} : p \times q$

iii) la fonctionnelle statistique associée à la matrice de corrélation en G est notée

$$R(G) = \begin{pmatrix} R_{YY}(G) & R_{YX}(G) \\ R_{XY}(G) & R_{XX}(G) \end{pmatrix}$$

avec $R_{YY}(G) : p \times p,\ R_{XX}(G) : q \times q,\ R_{YX}(G) = R'_{XY}(G) : p \times q$
et si on pose $D_Y(G) = \mathrm{diag}(C_{YY}(G))$ et $D_X(G) = \mathrm{diag}(C_{XX}(G))$ alors

$$\begin{aligned} R_{YY}(G) &= D_Y^{-1/2}(G)C_{YY}(G)D_Y^{-1/2}(G) \\ R_{XX}(G) &= D_X^{-1/2}(G)C_{XX}(G)D_X^{-1/2}(G) \\ R_{XY}(G) &= D_X^{-1/2}(G)C_{XY}(G)D_Y^{-1/2}(G) = R'_{YX}(G). \end{aligned}$$

Donc R est l'estimateur de la matrice de corrélation déduit de l'estimateur C affin équivariant de la matrice de covariance.

Par suite nous avons

$$R(F) = P, \quad D_X(F) = \Sigma_{DX}, \quad D_Y(F) = \Sigma_{DY}.$$

Pour compléter la notation posons

$$\begin{aligned} A^R &= R_{XX}^{-1}(G)R_{XY}(G)R_{YX}(G) \\ &= D_X^{1/2}(G)C_{XX}^{-1}(G)C_{XY}(G)D_Y^{-1}(G)C_{YX}(G)D_X^{-1/2}(G) \end{aligned}$$

et posons aussi

$$\mathcal{A}^R = A^R(F) = P_{XX}^{-1}P_{XY}P_{YX} = \Sigma_{DX}^{1/2}\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{DY}^{-1}\Sigma_{YX}\Sigma_{DX}^{-1/2}.$$

Enfin notons par $\lambda_{R,k}$ et $\alpha_{R,k}$ les valeurs et vecteurs propres de $\mathcal{A}^R$ avec $\ell_{R,k}(F) = \lambda_{R,k}$ et $a_{R,k}(F) = \alpha_{R,k}$.

On énonce un lemme et deux théorèmes dont les preuves se trouvent dans l'Annexe.

LEMME 4. – *Soit R un estimateur de la matrice de corrélation P déduit d'un estimateur C affin équivariant de la matrice de covariance Σ. Alors la fonction*

d'influence de $A^R = D_X^{1/2} C_{XX}^{-1} C_{XY} D_Y^{-1} C_{YX} D_X^{-1/2}$ *basée sur* R *en* $F = N(\mu, \Sigma)$ *est donnée par*

$$IF(Z, A^R, F) = \gamma_C(d(Z)) \big[P_{XX}^{-1} (\widetilde{X}\widetilde{Y}' P_{YX} + P_{XY}\widetilde{Y}\widetilde{X}' - \widetilde{X}\widetilde{X}' \mathcal{A}^R - P_{XY} G_{\widetilde{Y}} P_{YX}) + \frac{1}{2}(G_{\widetilde{X}} \mathcal{A}^R - \mathcal{A}^R G_{\widetilde{X}}) \big] \tag{29}$$

où pour tout vecteur W, $\widetilde{W} = \Sigma_{DW}^{-1/2}(W - \mu_W)$ *et* $G_{\widetilde{W}} = \mathrm{diag}(\widetilde{W}\widetilde{W}')$.

THÉORÈME 3. – *Sous les mêmes hypothèses, avec* $\widetilde{t}_{R,j} = \widetilde{\alpha}'_{R,j}\widetilde{X}$, *la fonction d'influence de* $\ell_{R,j}$ *est donnée par*

$$IF(Z, \ell_{R,j}, F) = \gamma_C(d(Z)) \big[2\widetilde{Y}' E_j \widetilde{t}_{R,j} - \lambda_{R,j} \widetilde{t}_{R,j}^2 - E_j' G_{\widetilde{Y}} E_j \big] \tag{30}$$

où $E_j = P_{YX}\alpha_{R,j}$ *est le vecteur des coefficients de régression de* $\widetilde{Y}$ *sur* $\widetilde{t}_{R,j}$.

Remarque. – On voit bien encore que

$$IF(Z, \ell_{R,j}, F) = \gamma_C(d(Z)) IF(Z, \ell_{R,j}^{\mathrm{cov}}, F). \tag{31}$$

THÉORÈME 4. – *Sous les mêmes hypothèses,*

$$\begin{aligned} IF(Z, a_{R,j}, F) = \gamma_C(d(Z)) \sum_{\substack{k=1 \\ k \neq j}}^{s} \frac{1}{\lambda_{R,j} - \lambda_{R,k}} & \Big[\widetilde{Y}' E_j \widetilde{t}_{R,k} + \widetilde{Y}' E_k \widetilde{t}_{R,j} - \lambda_{R,j} \widetilde{t}_{R,k} \widetilde{t}_{R,j} \\ & - E_k' G_{\widetilde{Y}} E_j + \frac{1}{2}(\lambda_{R,j} - \lambda_{R,k})(\alpha'_{R,k} P_{XX} G_{\widetilde{X}} \alpha_{R,j}) \Big] \alpha_{R,k} \\ & - \frac{1}{2}[\gamma_c(d(Z)) t_{R,j}^2 - \delta_c(d(Z))] \alpha_{R,j}. \end{aligned} \tag{32}$$

Remarque. – La fonction d'influence peut être utilisée pour spécifier la loi asymptotique d'un estimateur T_n d'une fonctionnelle T assez régulière si n est suffisamment grand. En effet, si l'on a

$$\sqrt{n}(T_n - T(F)) \xrightarrow{\mathcal{L}} N(0, \mathrm{VAS}(T, F)) \tag{33}$$

où VAS(T, F) est la matrice de covariance asymptotique de T_n, alors VAS(T, F) est donnée par

$$\text{VAS}(T, F) = E_F(IF(Z, T, F))(IF(Z, T, F)'). \tag{34}$$

La fonctionnelle T peut être n'importe lequel des paramètres estimés précédents $\widehat{\lambda}_j = \ell_j(F_n)$, $\widehat{\alpha}_j = a_j(F_n)$, $\widehat{\lambda}_{R,j} = \ell_{R,j}(F_n)$ ou $\widehat{\alpha}_{R,j} = a_{R,j}(F_n)$. Considérons par exemple $\widehat{\lambda}_j$. On sait par Lopuhaä (2000) que l'estimateur RMCD de la matrice de covariance est asymptotiquement normalement distribué. Donc ℓ_j et a_j le sont également et en particulier nous pouvons écrire

$$\sqrt{n}(\widehat{\lambda}_j - \lambda_j) \xrightarrow{\mathcal{L}} N(0, \text{VAS}(\widehat{\lambda}_j)) \tag{35}$$

où

$$\text{VAS}(\widehat{\lambda}_j) = E_F(IF^2(Z, \ell_j, F)) \tag{36}$$

et l'on peut approximer var$(\widehat{\lambda}_j)$ par

$$\text{var}(\widehat{\lambda}_j) \approx \frac{1}{n^2} \sum_{i=1}^{n} IF(Z_i, \ell_j, \widehat{F}_n)^2,\ 1 \leqslant j \leqslant s \tag{37}$$

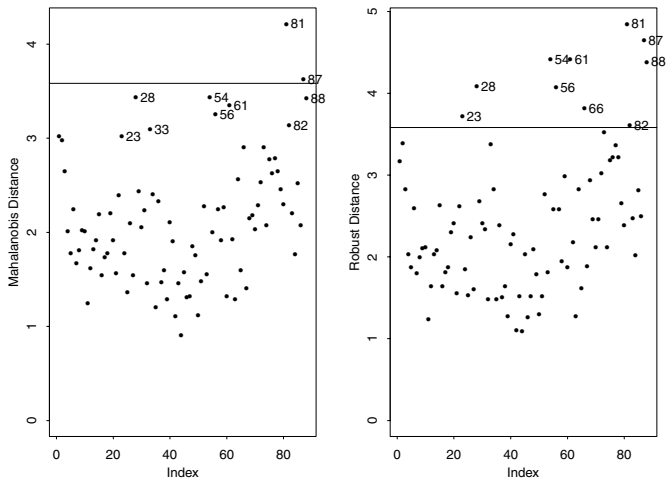
où $\widehat{F}_n$ est la fonction de répartition de la loi $N(T_n, C_n)$ et où T_n et C_n sont les estimateurs de location et de dispersion multivariées.

Cela peut permettre, entre autres choses, comme dans Lazraq et Cléroux (2002 A) pour le cas classique, de construire un algorithme pas à pas de sélection des variables latentes en analyse de la redondance robuste si n est suffisamment grand. Il faut de toute façon que $m < n(1 - \alpha)$, où α est le point de rupture, pour que l'estimateur RMCD soit possible.

7. Un exemple

Considérons les données (Mardia *et al* (1979) pp. 3–4) sur les notes obtenues par 88 étudiants lors de deux types d'examens : examens à livres fermés sur deux matières (mécanique et calcul vectoriel) et examens à livres ouverts sur trois matières (algèbre, analyse et statistique). Notons par Y_1 et Y_2 les résultats obtenus aux examens de mécanique et de calcul vectoriel respectivement et par X_1, X_2 et X_3 ceux obtenus aux examens d'algèbre, d'analyse et de statistique respectivement. On étudie la relation entre $Y = (Y_1, Y_2)'$ et $X = (X_1, X_2, X_3)'$ par l'analyse de la redondance. On a donc $n = 88$, $p = 2$ et $q = 3$

On calcule d'abord les distances de Mahalanobis de chaque observations en se basant d'une part sur l'estimateur classique de la matrice de covariance et d'autre part sur son estimateur robuste. Dans ce dernier cas nous avons utilisé l'algorithme RMCD du logiciel SPlus6 de la Société Mathsoff (2000) avec un point de rupture $\alpha = 10\%$ et paramètre $\delta = 0,025$. Le Graphique 4 montre les résultats. On remarque



GRAPHIQUE 4
Distances de Mahalanobis basées sur l'estimateur classique (à gauche) de la matrice de covariance et sur l'estimateur robuste (à droite)

qu'en se basant sur l'estimateur classique on identifie deux observations (81 et 87) douteuses tandis que selon l'approche robuste on en identifie dix (23, 82, 66, 28, 56, 54, 61, 88, 87 et 81). Dans ce dernier cas, on doit remettre en cause un plus grand nombre de données avant de poursuivre l'étude. On procède ensuite à une analyse de la redondance de Y sur X en utilisant chacun des deux estimateurs. Les résultats sont au Tableau 1.

TABLEAU 1
Résultats de l'analyse de la redondance de Y sur X

	Estimateur classique	Estimateur robuste
$RI(Y, t_1)$	$0,328$	$0,372$
$RI(Y, t_2)$	$3,596 \times 10^{-4}$	$3,363 \times 10^{-4}$
$\widehat{\lambda}_1$	$157,006$	$140,690$
(écart-type)	$(47,67)$	$(52,74)$
$\widehat{\lambda}_2$	$0,172$	$0,127$
(écart-type)	$(1,000)$	$(0,908)$
$\widehat{\alpha}_1$	$(0,083, 0,007, 0,004)$	$(0,084, 0,0200, 0,007)$
(écarts-type)	$((0,084), (0,092), (0,013))$	$((0,094), (0,185), (0,087))$
$\widehat{\alpha}_2$	$(-0,089, 0,099, -0,014)$	$(-0,054, 0,109, -0,051)$
(écarts-type)	$((0,074), (0,010), (0,003))$	$((0,131), (0,035), (0,010))$

Dans les deux cas $RI(Y,t_1)/RI(Y,X) \simeq 1$, ce qui signifie qu'une seule composante réussit à expliquer Y aussi bien que le vecteur X. A partir des écarts-type on réalise que les intervalles de confiance associés à λ_2 permettent, dans les deux cas, d'accepter la nullité de la valeur propre λ_2 confirmant ainsi une conclusion déjà obtenue dans Lazraq et Cléroux (2002A).

8. Simulations

Dans cette section on compare par simulation la qualité des estimateurs RI, $\widehat{\lambda}_k$ et $\widehat{\alpha}_k$ calculés à partir d'une matrice de dispersion C_n robuste et à partir de la matrice de covariance empirique classique S_n. On procède selon le même plan de simulation que dans Lazraq et Cléroux (2002 B) c'est-à-dire que l'on simule des données vectorielles à partir des paramètres suivants :

(i) $p = 2$ et $q = 3$

(ii) $n = 50,\ 75$ et 100

(iii) $L = 100$ échantillons générés pour chaque cas

(iv) deux types de lois sont considérées

a) données non contaminées : on génère les données à partir de la loi $N_{p+q}(0,\Sigma)$ avec $\Sigma = \begin{pmatrix} \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_{XX} \end{pmatrix}$ où $\Sigma_{YY} = I_p$, $\Sigma_{XX} = I_q$ et $\Sigma_{YX} = \Sigma'_{XY} = C_{10}$, C_{20} et C_{30} où C_{xy} a tous ses éléments égaux à $0 \cdot xy$

b) données contaminées : pour $n_C = \lfloor 0.1n \rfloor$ (partie entière de 10% de n) on génère $n - n_C$ données à partir de la loi multinormale précédente et n_C données à partir de la loi $N_{p+q}(0, 49I)$.

Dans chaque cas, pour chacune des $L = 100$ simulations, $\ell = 1, 2, \ldots, L$, on calcule

1) ρI, $RI^{(\ell)}(S_n)$ et $RI^{(\ell)}(C_n)$. Par analogie à Lazraq et Cléroux (2002 B) et Dehon *et al* (2000) on stabilise la variance par la transformation $W(RI^{(\ell)}) = tanh^{-1}(RI^{(\ell)})$ appliquée successivement à ρI, $RI^{(\ell)}(S_n)$ et $RI^{(\ell)}(C_n)$ puis on calcule la moyenne des carrés des erreurs d'estimation $e^2_{RI} = MCE(RI) = \frac{1}{L}\sum_{\ell=1}^{L}(W(RI^{(\ell)}) - W(\rho I))^2$ avec $RI^{(\ell)}(S_n)$ et $RI^{(\ell)}(C_n)$ pour obtenir respectivement $e^2_{RI(S_n)}$ et $e^2_{RI(C_n)}$.

2) λ (la plus grande valeur propre), $\lambda^{(\ell)}(S_n)$ et $\lambda^{(\ell)}(C_n)$ et par analogie à Croux et Hoesbroeck (1999) on prend la ransformation logarithmique pour calculer $e^2_\lambda = MCE(\widehat{\lambda}) = \frac{1}{L}\sum_{\ell=1}^{L}(\log\widehat{\lambda}^{(\ell)} - \log\lambda)^2$ avec $\widehat{\lambda}^{(\ell)}(S_n)$ et $\widehat{\lambda}^{(\ell)}(C_n)$. On obtient alors $e^2_{(\lambda)(S_n)}$ et $e^2_{(\lambda)(C_n)}$.

3) le vecteur propre α associé à λ ainsi que $\widehat{\alpha}^{(\ell)}(S_n)$ et $\widehat{\alpha}^{(\ell)}(C_n)$ et par analogie à Dehon *et al* (2000) on utilise pour $MCE(\widehat{\alpha})$ la somme $e^2_\alpha = MCE(\widehat{\alpha}) = \frac{1}{L}\sum_{\ell=1}^{L}\cos^{-1}\left(\frac{\| \alpha'\widehat{\alpha}^{(\ell)} \|}{\| \alpha \| \| \widehat{\alpha}^{(\ell)} \|}\right)$. Cette mesure est la valeur moyenne des angles

positifs entre $\widehat{\alpha}^{(\ell)}$ et α et est invariante par rapport au choix de la constante de normalisation pour l'analyse de la redondance.

TABLEAU 2
Sommes des carrés des erreurs d'estimation pour les données non contaminées

n	mesure	C_{10}	C_{20}	C_{30}
50	$e^2_{RI(S_n)}$	$0,015$	$0,012$	$0,008$
	$e^2_{RI(C_n)}$	$0,100$	$0,065$	$0,033$
	$e^2_{\lambda(S_n)}$	$0,976$	$0,279$	$0,105$
	$e^2_{\lambda(C_n)}$	$2,506$	$0,526$	$0,312$
	$e^2_{\alpha(S_n)}$	$0,758$	$0,324$	$0,185$
	$e^2_{\alpha(C_n)}$	$0,880$	$0,725$	$0,399$
75	$e^2_{RI(S_n)}$	$0,010$	$0,006$	$0,007$
	$e^2_{RI(C_n)}$	$0,038$	$0,030$	$0,020$
	$e^2_{\lambda(S_n)}$	$0,683$	$0,179$	$0,070$
	$e^2_{\lambda(C_n)}$	$1,285$	$0,387$	$0,199$
	$e^2_{\alpha(S_n)}$	$0,661$	$0,275$	$0,140$
	$e^2_{\alpha(C_n)}$	$0,854$	$0,544$	$0,306$
100	$e^2_{RI(S_n)}$	$0,005$	$0,005$	$0,005$
	$e^2_{RI(C_n)}$	$0,023$	$0,009$	$0,014$
	$e^2_{\lambda(S_n)}$	$0,484$	$0,146$	$0,065$
	$e^2_{\lambda(C_n)}$	$0,997$	$0,228$	$0,164$
	$e^2_{\alpha(S_n)}$	$0,536$	$0,235$	$0,130$
	$e^2_{\alpha(C_n)}$	$0,754$	$0,441$	$0,206$

Remarques. –

(i) Dans les calculs précédents et les tables qui suivent, nous ne nous intéressons qu'à une seule valeur propre et un seul vecteur propre pour la raison suivante :

TABLEAU 3
Sommes des carrés des erreurs d'estimation pour les données contaminées

n	mesure	C_{10}	C_{20}	C_{30}
50	$e^2_{RI(S_n)}$	$0,205$	$0,093$	$0,054$
	$e^2_{RI(C_n)}$	$0,089$	$0,040$	$0,035$
	$e^2_{\lambda(S_n)}$	$14,318$	$6,564$	$3,564$
	$e^2_{\lambda(C_n)}$	$2,396$	$0,410$	$0,261$
	$e^2_{\alpha(S_n)}$	$0,995$	$0,956$	$1,018$
	$e^2_{\alpha(C_n)}$	$0,905$	$0,632$	$0,316$
75	$e^2_{RI(S_n)}$	$0,132$	$0,049$	$0,025$
	$e^2_{RI(C_n)}$	$0,033$	$0,025$	$0,018$
	$e^2_{\lambda(S_n)}$	$12,690$	$5,328$	$2,068$
	$e^2_{\lambda(C_n)}$	$1,191$	$0,408$	$0,162$
	$e^2_{\alpha(S_n)}$	$1,031$	$0,973$	$0,996$
	$e^2_{\alpha(C_n)}$	$0,891$	$0,564$	$0,273$
100	$e^2_{RI(S_n)}$	$0,072$	$0,017$	$0,037$
	$e^2_{RI(C_n)}$	$0,018$	$0,009$	$0,013$
	$e^2_{\lambda(S_n)}$	$10,965$	$3,562$	$1,473$
	$e^2_{\lambda(C_n)}$	$0,801$	$0,169$	$0,171$
	$e^2_{\alpha(S_n)}$	$0,974$	$1,000$	$0,957$
	$e^2_{\alpha(C_n)}$	$0,706$	$0,382$	$0,197$

dans tous les cas de figure (C_{10}, C_{20}, C_{30}) nous avons calculé le rapport $\dfrac{\rho I(Y,t_1)}{\rho I(Y,X)}$ où $t_1 = \alpha' X$ et où α est le premier vecteur propre. Il arrive que ce rapport est approximativement égal à 1 dans tous les cas et que la première composante t_1 de l'analyse de la redondance explique aussi bien Y que ne le ferait le vecteur X.

(ii) Nous avons utilisé l'algorithme $RMCD$ du logiciel SPlus6 avec un point de rupture de 50% et paramètre $\delta = 0.025$.

Les résultats se trouvent aux Tableaux 2 et 3 et on peut facilement en tirer les conclusions suivantes :

a) les estimateurs classiques sont supérieurs aux estimateurs robustes lorsque les données ne sont pas contaminées. En effet on remarque au Tableau 2 que $e_U^2(S_n) < e_U^2(C_n)$ pour tout $U \in \{RI, \lambda, \alpha\}$ dans tous les cas.

b) les estimateurs robustes sont supérieurs aux estimateurs classiques lorsque les données sont contaminées puisque $e_U^2(C_n) < e_U^2(S_n)$ pour tout $U \in \{RI, \lambda, \alpha\}$ dans tous les cas.

c) de façon générale et comme prévu la qualité des estimateurs s'améliore lorsque n et/ou xy augmente. Cependant, dans le cas classique (Tableau 3), il arrive que certaines erreurs moyennes augmentent avec n. Cela s'explique par le fait que dans le plan de simulation, le nombre de données contaminées augmente aussi avec n.

9. Conclusion

Dans cet article nous avons étudié l'analyse de la redondance robuste basée soit sur la matrice de covariance, soit sur la matrice de corrélation. En utilisant des estimateurs $RMCD$ nous avons obtenu les paramètres de l'analyse de la redondance ainsi que leurs fonctions d'influence. Un exemple et des simulations montrent clairement l'avantage de l'approche robuste en présence de données contaminées.

Remerciements

Les auteurs reconnaissent la grande compétence de l'arbitre et le remercient pour son excellent travail. Ils remercient également le CRSNG qui a en partie financé cette recherche.

Annexe

Dans cette annexe on démontre les résultats de la section 6.

Preuve du Lemme 4. – Nous avons $A^R = D_X^{1/2} C_{XX}^{-1} C_{XY} D_Y^{-1} C_{YX} D_X^{-1/2}$ d'où on tire

$$\begin{aligned} IF(Z, A^R, F) = {} & IF(Z, D_X^{1/2}, F)\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{DY}^{-1}\Sigma_{YX}\Sigma_{DX}^{-1/2} \\ & + \Sigma_{DX}^{1/2} IF(Z, C_{XX}^{-1}, F)\Sigma_{XY}\Sigma_{DY}^{-1}\Sigma_{YX}\Sigma_{DX}^{-1/2} \\ & + \Sigma_{DX}^{1/2}\Sigma_{XX}^{-1} IF(Z, C_{XY}, F)\Sigma_{DY}^{-1}\Sigma_{YX}\Sigma_{DX}^{-1/2} \\ & + \Sigma_{DX}^{1/2}\Sigma_{XX}^{-1}\Sigma_{XY} IF(Z, D_Y^{-1}, F)\Sigma_{YX}\Sigma_{DX}^{-1/2} \\ & + \Sigma_{DX}^{1/2}\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{DY}^{-1} IF(Z, C_{YX}, F)\Sigma_{DX}^{-1/2} \\ & + \Sigma_{DX}^{1/2}\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{DY}^{-1}\Sigma_{YX} IF(Z, D_X^{-1/2}, F). \end{aligned} \tag{38}$$

Étudions séparément chaque terme du membre de droite de (38).

a) **Premier terme**

Puisque $D_X^{1/2} D_X^{1/2} = D_X$ on peut écrire

$$IF(Z, D_X^{1/2}, F)\Sigma_{DX}^{1/2} + \Sigma_{DX}^{1/2} IF(Z, D_X^{1/2}, F) = IF(Z, D_X, F) \tag{39}$$

et comme Σ_{DX} est une matrice diagonale, (39) devient

$$\begin{aligned} IF(Z, D_X^{1/2}, F) &= \frac{1}{2}\Sigma_{DX}^{-1/2} IF(Z, D_X, F) \\ &= \frac{1}{2}\Sigma_{DX}^{-1/2}[\gamma(d(Z))diag(X-\mu_X)(X-\mu_X)' - \delta(d(z))\Sigma_{DX}] \\ &= \frac{1}{2}[\gamma(d(Z))G_{\widetilde{X}} - \delta(d(Z))I_q]\Sigma_{DX}^{1/2} \end{aligned}$$

par (13) et puisque $X - \mu_X = \Sigma_{DX}^{1/2}\widetilde{X}$. Le premier terme de (38) s'écrit donc

$$\frac{1}{2}[\gamma(d(Z))G_{\widetilde{X}} - \delta(d(Z))I_q]P_{XX}^{-1}P_{XY}P_{YX}. \tag{40}$$

b) **Second terme**

À partir de (18) et en utilisant encore une fois les définitions des sous-matrices de la matrice P, le second terme s'écrit

$$-P_{XX}^{-1}[\gamma_C(d(Z))\widetilde{X}\widetilde{X}' - \delta_C(d(Z))P_{XX}]P_{XX}^{-1}P_{XY}P_{YX}. \tag{41}$$

c) **Troisième terme**

Par (13) on a

$$IF(Z, C_{XY}, F) = \gamma(d(Z))(X - \mu_X)(Y - \mu_Y)' - \delta(d(Z))\Sigma_{XY}$$

et le troisième terme s'écrit

$$P_{XX}^{-1}[\gamma_C(d(Z))\widetilde{X}\widetilde{Y}' - \delta_c(d(Z))P_{XY}]P_{YX}. \tag{42}$$

d) **Quatrième et cinquième termes**

On procède comme précédemment pour obtenir respectivement

$$-P_{XX}^{-1}P_{XY}[\gamma_c(d(Z))G_{\widetilde{Y}} - \delta_C(d(Z))I_p]P_{YX} \tag{43}$$

$$P_{XX}^{-1}P_{XY}[\gamma_c(d(Z))\widetilde{Y}\widetilde{X}' - \delta_C(d(Z))P_{YX}]. \tag{44}$$

e) **Sixième terme**

On réalise d'abord que

$$IF(Z, D_X^{-1/2}, F) = -\Sigma_{DX}^{-1/2} IF(Z, D_X^{1/2}, F)\Sigma_{DX}^{-1/2} = -\frac{1}{2}\Sigma_{DX}^{-1} IF(Z, D_X, F)\Sigma_{DX}^{-1/2}$$

et le sixième terme s'écrit

$$-\frac{1}{2}P_{XX}^{-1}P_{XY}P_{YX}[\gamma_C(d(Z))G_{\widetilde{X}} - \delta_C(d(Z))I_q]. \tag{45}$$

En reportant (40) à (45) dans (38), les termes en $\delta(d(Z))$ s'annulent et on obtient (29), démontrant le Lemme 4. □

Preuve du Théorème 3. – En combinant (14) et (29) on obtient

$$\begin{aligned} IF(Z, \ell_{R,j}, F) &= \alpha'_{R,j}P_{XX}IF(Z, A^R, F)\alpha_{R,j} \\ &= \gamma_C(d(Z))\big[\widetilde{t}_{R,j}\widetilde{Y}'E_j + E_j'\widetilde{Y}t_{R,j} - \lambda_{R,j}\widetilde{t}_{R,j}^2 - E_j'G_{\widetilde{Y}}E_j \\ &+ \frac{1}{2}(\lambda_{R,j}\alpha'_{R,j}P_{XX}G_{\widetilde{X}}\alpha_{R,j} - \alpha'_{R,j}P_{XY}P_{YX}G_{\widetilde{X}}\alpha_{R,j})\big] \end{aligned} \tag{46}$$

Or nous avons $P_{XY}P_{YX}\alpha_{R,j} = \lambda_{R,j}P_{XX}\alpha_{R,j}$ de sorte que la parenthèse s'annule et on obtient le résultat désiré. □

Preuve du Théorème 4. – À partir de (15) on peut écrire

$$IF(Z,a_{R,j},F)=\sum_{\substack{k=1\\k\neq j}}^{s}\frac{\alpha'_{R,k}P_{XX}IF(Z,A^R,F)\alpha_{R,j}}{\lambda_{R,j}-\lambda_{R,k}}\alpha_{R,k}$$
$$-\frac{1}{2}(\alpha'_{R,j}IF(Z,R_{XX},F)\alpha_{R,j})\alpha_{R,j}. \qquad (47)$$

Or, par (29) nous avons

$$\begin{aligned}\alpha'_{R,k}P_{XX}IF(Z,A^R,F)\alpha_{R,j}&=\gamma_C(d(Z))[\alpha'_{R,k}\widetilde{X}\widetilde{Y}'P_{YX}\alpha_{R,j}\\&+\alpha'_{R,k}P_{XY}\widetilde{Y}\widetilde{X}'\alpha_{R,j}-\alpha'_{R,k}\widetilde{X}\widetilde{X}'\mathcal{A}^R\alpha_{R,j}\\&-\alpha'_{R,k}P_{XY}G_{\widetilde{Y}}P_{YX}\alpha_{R,j}\\&+\frac{1}{2}(\alpha'_{R,k}P_{XX}G_{\widehat{X}}\mathcal{A}^R\alpha_{R,j}\\&-\alpha'_{R,k}P_{XX}\mathcal{A}^RG_{\widetilde{X}}\alpha_{R,j})] \qquad (48)\end{aligned}$$

En notant que $P_{XX}\mathcal{A}^R=P_{XY}P_{YX}$ et $\alpha'_{R,k}P_{XY}P_{YX}=\lambda_{R,k}\alpha'_{R,k}P_{XX}$ on écrit

$$\begin{aligned}\alpha'_{R,k}P_{XX}IF(Z,A^R,F)\alpha_{R,j}&=\gamma_C(d(Z)[\widetilde{Y}'E_j\widetilde{t}_{R,k}+\widetilde{Y}'E_k\widetilde{t}_{R,j}-\lambda_{R,j}\widetilde{t}_{R,k}\widetilde{t}_{R,j}\\&-E'_kG_{\widetilde{Y}}E_j+\frac{1}{2}(\lambda_{R,j}-\lambda_{R,k})(\alpha'_{R,k}P_{XX}G_{\widetilde{X}}\alpha_{R,j})]. \qquad (49)\end{aligned}$$

D'autre part, compte tenu de ce que $\alpha'_{R,j}P_{XX}\alpha_{R,j}=1$, on a

$$\begin{aligned}\alpha'_{R,j}IF(Z,R_{XX},F)\alpha_{R,j}&=\gamma_C(d(Z))[\alpha'_{R,j}(\widetilde{X}\widetilde{X}')\alpha_{R,j}-\delta_C(d(Z))]\\&=\gamma_C(d(Z))t^2_{R,j}-\delta_C(d(Z)). \qquad (50)\end{aligned}$$

Et en reportant (49) et (50) dans (47) on obtient la formule (32) du Théorème 4. □

Références

BRY X. (1996), Analyse factorielles multiples, Paris, France, *Economica*.

CRAMER E.M. and NICEWANDER W.A. (1979), Some Symetric, Invariant Measures of Multivariate Association, *Psychometrika*, **44**, 43–54.

CROUX C. et DEHON C. (2002), Analyse canonique basée sur des estimateurs robustes de la matrice de covariance, *Rev. Statistique Appliquée*, **L(2)**, 5–26.

CROUX C. and HAESBROECK G. (2000), Principal Component Analysis Based on Robust Estimators of the Covariance or Correlation Matrix : Influence Functions and Efficiencies, *Biometrika*, **87,3**, 603–618.

CROUX C. and HAESBROECK G. (1999), Influence Function and Efficiency of the Minimum Covariance Determinant Scatter Matrix Estimator, *J. Multivariate Analysis*, **71**, 161–190.

DEHON C., FITZMOSER P. and CROUX C. (2000), Robust Methods for Canonical Correlation Analysis, in *Data Analysis, Classification and Related Methods*, Eds. H.A.L. Kiers, J.P. Rosson, P.J.E. Groenen and M. Schader, Springer, Berlin, pp. 321–326.

GLEASON T.C. (1976), On Redundancy in Canonical Correlation, *Psychological Bulletin*, **83**, 1004–1006.

HAMPEL F.R. (1974), The Influence Curve and its Role in Robust Estimation, *J. American Statistical Association*, **69**, 383–393.

LAZRAQ A. et CLÉROUX R. (1988), Étude comparative de différentes mesures de liaison entre deux vecteurs aléatoires, *Statistique et Analyse des données*, **13**, 39–58.

LAZRAQ A. et CLÉROUX R. (2002A), Testing the Significance of the Successive Components in Redundancy Analysis, *Psychometrika*, **67**, 411–419.

LAZRAQ A. et CLÉROUX R. (2002B), Inférence robuste sur un indice de redondance, *Rev. Statistique Appliquée*, **L(4)**, 39–54.

LOPUHAÄ H.P. (2000), Asymptotics of Reweighted Estimators of Multivariate Location and Scatter, *The Annals of Statistics*, **27**, 1638–1665.

MARDIA K.V., KENT J.T. and BIBBY J.M. (1979), *Multivariate Analysis*, London, U.K., Academic Press.

RAO C.R. (1964), The Use and Interpretation of Principal Component Analysis in Applied Research, *Sankhya*, **26**, 329–358.

RAO C.R. (1973), *Linear Statistical Inference and its Applications*, 2^{nd} ed., New York, John Wiley and sons.

ROUSSEEUW P.J (1985), Multivariate Estimation with High Breakdown Point. *Mathematical Statistics and Applications*, **vol. B.**, pp. 283–297, Dordrecht, The Netherland, W. Grossman, G. Pflug, I. Vincze and W. Wertz, eds.

SPLUS6 User's Guide (2000), MathSoft Inc., Seattle.

STEWARD D. and LOVE W. (1968), A General Canonical Corelation Index, *Psychological Bulletin*, **70**, 160-163.

VAN DEN WOLLENBERG A.L. (1977), Redundancy Analysis, an Alternative for Canonical Correlation Analysis, *Psychometrika*, **42**, 207–219.