REVUE DE STATISTIQUE APPLIQUÉE

CHRISTOPHE CROUX
CATHERINE DEHON

Analyse canonique basée sur des estimateurs robustes de la matrice de covariance

Revue de statistique appliquée, tome 50, n° 2 (2002), p. 5-26
<http://www.numdam.org/item?id=RSA_2002__50_2_5_0>

Rev. Statistique Appliquée, 2002, **L** (2), 5-26

ANALYSE CANONIQUE BASÉE SUR DES ESTIMATEURS ROBUSTES DE LA MATRICE DE COVARIANCE

Christophe CROUX et Catherine DEHON *

Université Libre de Bruxelles

RÉSUMÉ

Dans ce papier, nous proposons une approche robuste pour l'analyse canonique, qui est une méthode majeure de l'analyse multivariée. L'analyse canonique s'appuie sur l'estimation de la matrice de covariance. L'estimateur classique, le plus puissant sous l'hypothèse de normalité, devient complètement non fiable lorsque les données contiennent des points aberrants. Nous proposons une solution à ce problème en utilisant un estimateur robuste de la matrice de covariance. L'estimateur robuste utilisé est le «Minimum Covariance Determinant» estimateur, qui a un haut point de rupture. Nous donnons également une expression des fonctions d'influence des corrélations et vecteurs canoniques dans le cas général où la matrice de covariance est estimée par n'importe quel estimateur régulier multivarié de dispersion. Nous traitons aussi de manière robuste une base de données réelles.

Mots-clés : *Corrélation canonique, Fonction d'influence, Robustesse.*

ABSTRACT

In this paper we present a robust approach to canonical correlation analysis, one of the major methods in multivariate statistics. A canonical correlation analysis is based on estimators of covariance matrices. The classical estimator, most powerful under the assumption of normality, becomes unreliable in presence of outlying observations. A remedy for this problem is to use robust estimators of the covariance matrix. Here we will use the "Minimum Covariance Determinant," estimator, which has a high breakdown point. We also give an expression for the influence function of estimators of the canonical correlations and vectors when using any regular robust estimator of the dispersion matrix. A real data example will be given.

Keywords : *Canonical correlation, Influence function, Robustness.*

1. Introduction et motivation

L'analyse canonique a été introduite par Harrold Hotelling (1936). Elle joue un rôle théorique important puisqu'elle permet de lier les formalismes de plusieurs

* ECARES, Faculté SOCO et Institut de Statistique, Université Libre de Bruxelles, CP-114, av. F.D. Roosevelt 50, B-1050 Brussels, Belgium.

techniques de la statistique multivariée. Son objectif est de caractériser les relations linéaires existantes entre deux ensembles de variables aléatoires. Pour introduire l'analyse canonique, considérons deux ensembles de variables. Le premier groupe, que nous noterons par le vecteur aléatoire X, sera composé de p variables et le deuxième groupe, noté Y, sera de taille q. Les vecteurs aléatoires X et Y ont donc respectivement les dimensions $(p,1)$ et $(q,1)$. Sans perte de généralité, nous supposons que $p \leqslant q$. Nous avons au niveau des deux populations considérées

$$\begin{aligned} E(X) = \mu_X \text{ et } E(Y) = \mu_Y, \\ \mathrm{Cov}(X) = \Sigma_{XX} \text{ et } \mathrm{Cov}(Y) = \Sigma_{YY}, \\ \mathrm{Cov}(X,Y) = \Sigma_{XY} \text{ et } \mathrm{Cov}(Y,X) = \Sigma_{YX}. \end{aligned}$$

Nous supposons que toutes les matrices considérées sont de rang maximum. Les vecteurs aléatoires X et Y peuvent être considérés conjointement sous la forme d'une variable aléatoire Z, de dimension $d = p + q$. Le vecteur aléatoire Z s'écrit

$$Z = [X_1, \ldots, X_p, Y_1, \ldots, Y_q]^t,$$

le vecteur des moyennes populations est le suivant

$$\mu = \begin{bmatrix} E(X) \\ E(Y) \end{bmatrix} = \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix},$$

et la matrice de covariance

$$\Sigma = E(Z-\mu)(Z-\mu)^t = \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix}.$$

Le but de l'analyse canonique est de résumer, de la manière la plus adéquate, les relations entre les ensembles X et Y. La matrice de covariance nous fournit pq covariances, avec lesquelles il est parfois difficile de se faire une idée générale si les dimensions de p et/ou de q sont grandes. L'analyse canonique va nous permettre d'isoler l'information dans seulement quelques covariances bien choisies. Le problème sera de trouver les composantes canoniques qui seront des combinaisons linéaires déterminées dans chacun des deux groupes de variables et, les corrélations canoniques qui mesureront la corrélation entre les composantes canoniques. Plus précisement, la première étape déterminera la paire de combinaisons linéaires ayant la corrélation maximum. Dans la seconde étape, nous chercherons la paire de combinaisons linéaires ayant une corrélation maximum parmi toutes les paires non corrélées avec celle sélectionnée à la première étape. Et ce processus sera répété p fois pour obtenir p paires de composantes canoniques et p corrélations canoniques.

Les vecteurs $\alpha_1 \in I\!R^p$ et $\beta_1 \in I\!R^q$ tels que

$$(\alpha_1, \beta_1) = \underset{\alpha,\beta}{\mathrm{argmax}}\, \mathrm{Corr}(\alpha^t X, \beta^t Y) = \underset{\alpha,\beta}{\mathrm{argmax}}\, \frac{\alpha^t \Sigma_{XY} \beta}{\sqrt{(\alpha^t \Sigma_{XX} \alpha)(\beta^t \Sigma_{YY} \beta)}} \qquad (1)$$

sont alors définis comme la première paire de vecteurs canoniques. Les variables univariées résultantes

$$U_1 = \alpha_1^t X \text{ et } V_1 = \beta_1^t Y$$

sont appelées variables canoniques. La première corrélation canonique ρ_1 est définie par la valeur absolue de la corrélation entre les 2 variables canoniques, qui est donc le maximum atteint en (1). Les k-ièmes vecteurs canoniques (α_k, β_k), pour $2 \leqslant k \leqslant p$, sont définis par

$$(\alpha_k, \beta_k) = \underset{\alpha,\beta}{\operatorname{argmax}} \operatorname{Corr}(\alpha^t X, \beta^t Y)$$

sous les contraintes $\operatorname{Cov}(U_k, U_l) = 0, \operatorname{Cov}(V_k, V_l) = 0, \operatorname{Cov}(U_k, V_l) = 0$ et $\operatorname{Cov}(V_k, U_l) = 0$ où $l \in \{1, \ldots, k-1\}$ et $U_j = \alpha_j^t X$ et $V_j = \beta_j^t Y$ sont les j-ièmes variables canoniques $(2 \leqslant j \leqslant k)$. Les vecteurs canoniques étant uniques à un facteur multiplicatif près, on décide pour avoir unicité que les variables canoniques auront une variance unitaire

$$\begin{aligned} s_{U_j}^2 &= \operatorname{Var}(U_j) = \alpha_j^t \Sigma_{XX} \alpha_j = 1 \\ s_{V_j}^2 &= \operatorname{Var}(V_j) = \beta_j^t \Sigma_{YY} \beta_j = 1, \end{aligned} \tag{2}$$

pour $j = 1, \ldots, p$. Les contraintes d'orthogonalité et de normalisation sur les vecteurs canoniques, seront donc les suivantes : $\alpha_i^t \Sigma_{XX} \alpha_j = \delta_{ij}$, $\beta_i^t \Sigma_{YY} \beta_j = \delta_{ij}$ et $\alpha_i^t \Sigma_{XY} \beta_j = \rho_i \delta_{ij}$ où δ_{ij} est le delta de Kronecker.

Le calcul des corrélations et des vecteurs canoniques est basé sur une analyse de valeurs et vecteurs propres qui est la suivante : la première corrélation canonique ρ_1 est la racine carrée de la première (la plus grande) valeur propre de $\Sigma_{XX}^{-1} \Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{YX}$ ou de $\Sigma_{YY}^{-1} \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XY}$, et les vecteurs canoniques α_1 et β_1 sont les vecteurs propres correspondants. La k-ième corrélation canonique ρ_k est la racine carrée de la k-ième plus grande valeur propre de $\Sigma_{XX}^{-1} \Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{YX}$ ou de $\Sigma_{YY}^{-1} \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XY}$, et les vecteurs canoniques α_k et β_k sont les vecteurs propres correspondants où $k \leqslant p$. Pour un traitement plus détaillé de l'analyse canonique classique, voir Johnson & Wichern (1992, Chapitre 10) ou Rencher (1998, Chapitre 8).

Dans la Section 2, nous allons montrer un moyen d'obtenir des estimateurs robustes pour l'analyse canonique. Les estimateurs classiques pour les corrélations et vecteurs canoniques sont généralement obtenus en estimant les matrices de covariance population qui interviennent dans l'analyse par leurs contreparties empiriques. Nous allons robustifier cette procédure en utilisant des estimateurs robustes à la place des estimateurs échantillon. Les méthodes robustes proposées dans ce travail sont moins puissantes que la méthode classique dans la situation où toutes les hypothèses classiques sont vérifiées. Néanmoins, l'analyse canonique classique est tout à fait inappropriée si la base de données contient des points aberrants. Un utilisateur non averti pourra développer des interprétations à partir de résultats potentiellement erronés. Dans la Section 3, nous introduisons la fonctionnelle statistique qui correspondra aux estimateurs robustes. Ceci nous permettra de préparer l'obtention des fonctions d'influence de ces estimateurs, qui constitue la contribution principale de ce papier. Dans la Section 4, nous donnerons une expression des fonctions d'influence des

corrélations et vecteurs canoniques dans le cas général où la matrice de covariance est estimée par n'importe quel estimateur régulier de dispersion multivariée. Nous ferons ensuite le lien avec le cas classique. Un exemple réel sera analysé dans la Section 5, et nous finirons par une brève conclusion. Toutes les démonstrations seront reprises en annexe.

2. Estimation Robuste

Les corrélations canoniques et les vecteurs canoniques sont donc obtenus, comme vu précédemment, en calculant les valeurs et vecteurs propres de $\mathcal{M}_x$ et $\mathcal{M}_y$ où

$$\mathcal{M}_x = \Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX} \tag{3}$$
$$\mathcal{M}_y = \Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}. \tag{4}$$

Le calcul des corrélations et des vecteurs canoniques d'une base de données exigera donc l'estimation de la matrice de covariance Σ. La méthode classique utilise, pour un échantillon $Z_n = \{z_1, \ldots, z_n\}$ de n observations et de dimension d, la matrice de covariance échantillon

$$C_n = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})(z_i - \bar{z})^t \text{ où } \bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i.$$

Cet estimateur est optimal lorsque les données sont distribuées suivant une loi normale multivariée. Cependant il est extrêmement sensible face à des points aberrants, et dans ce cas, les conclusions faites sur la base des corrélations canoniques et des vecteurs canoniques peuvent être totalement erronées. Il est important de souligner que l'apparition de points aberrants est relativement fréquente dans des cas réels, surtout dans des bases de données de grande dimension. L'idée est donc de remplacer l'estimateur classique par un estimateur robuste de dispersion multivariée, pour obtenir une analyse canonique robuste. Pour chaque choix d'estimateur robuste de la matrice de covariance, une nouvelle méthode robuste d'analyse canonique sera obtenue ainsi que de nouveaux estimateurs robustes des corrélations canoniques et des vecteurs canoniques. Campbell (1982) utilise une combinaison de M-estimateur et des fonctions de poids basées sur des distances robustes, pour robustifier l'analyse canonique. Karnel (1991), lui, utilise comme estimateur multivarié, le M-estimateur. Nous allons travailler avec d'autres estimateurs que le M-estimateur car celui-ci n'est pas satisfaisant. En effet, plus la dimension augmente moins le M-estimateur est robuste (Maronna 1976). Son point de rupture, défini par la plus grande proportion de points aberrants dans les données ne faisant pas exploser ou imploser l'estimateur, est inférieur ou égal à $1/(d+1)$. Dans cet article, nous allons considérer deux estimateurs robustes à haut point de rupture, l'estimateur «Minimum Covariance Determinant» (MCD) et l'estimateur «Reweighted Minimum Covariance Determinant» (RMCD). Plus de détails concernant le point de rupture des estimateurs de la matrice de dispersion peuvent être trouvés dans Lopuhaä & Rousseeuw (1991).

L'estimateur MCD, introduit par Rousseeuw (1985), est déterminé pour un échantillon $\{z_1, \ldots, z_n\} \in I\!R^d$ en sélectionnant le sous échantillon $\{z_{i_1}, \ldots, z_{i_h}\}$ de taille h ($1 \leqslant h \leqslant n$), qui minimise le déterminant de la matrice de covariance parmi tous les sous-échantillons possibles de taille h. L'estimateur MCD de position est alors défini comme

$$T_n = \frac{1}{h} \sum_{k=1}^{h} z_{i_k}$$

et l'estimateur MCD de la matrice de covariance est défini par

$$C_n = c_1 \frac{1}{h} \sum_{k=1}^{h} (z_{i_k} - T_n)(z_{i_k} - T_n)^t$$

où $c_1 = (1-\alpha)/F_{\chi^2_{d+2}}(q_\alpha)$ ($F_{\chi^2_{d+2}}$ étant la fonction de répartition de la loi du khi-2 à $d+2$ degrés de liberté) est un facteur de convergence pour une distribution normale. Nous avons défini $q_\alpha = \chi^2_{d,1-\alpha}$ comme le quantile $(1-\alpha)$ de la distribution khi-2 à d degrés de liberté. Généralement, h prendra les valeurs $h = [n(1-\alpha)]$, avec $\alpha = 0.5$ ou $\alpha = 0.25$. Dans le premier cas, l'estimateur aura le plus haut point de rupture (50%) et dans le second cas on aura un meilleur compromis entre efficacité et haut point de rupture (qui sera de 25%). Des propriétés théoriques de l'estimateur MCD ont été investiguées par Butler, Davies & Jhun (1993). La fonction d'influence de l'estimateur MCD d'échelle multivariée a été trouvée par Croux & Haesbroeck (1999). Un algorithme rapide pour calculer l'estimateur MCD a été proposé par Rousseeuw & Van Driessen (1999).

Pour augmenter l'efficacité des estimateurs à haut point de rupture, il est souvent recommandé d'utiliser leur version pondérée en une étape, qui garde le point de rupture de l'estimateur initial. Une étude des propriétés asymptotiques d'un tel estimateur a été réalisée par Lopuhaä (2000) et les efficacités sont calculées par Croux & Haesbroeck (1999). Nous allons donc préférer utiliser les estimateurs RMCD (MCD pondérés à une étape) qui sont définis par

$$T_n^R = \frac{\sum_{i=1}^n w_i z_i}{\sum_{i=1}^n w_i}$$
$$C_n^R = c_2 \frac{\sum_{i=1}^n w_i (z_i - T_n^R)(z_i - T_n^R)^t}{\sum_{i=1}^n w_i}$$

où $c_2 = (1-\delta)/F_{\chi^2_{d+2}}(q_\delta)$ est un facteur de convergence. Les poids sont calculés comme suit

$$w_i = \begin{cases} 1 & \text{si } (z_i - T_n)^t C_n^{-1} (z_i - T_n) \leqslant q_\delta \\ 0 & \text{sinon} \end{cases}$$

où T_n et C_n sont les estimateurs MCD non pondérés. Dans notre travail, nous prendrons comme point de rupture 50%, c'est-à-dire $\alpha = 0.5$, et un paramètre

$\delta = 0.025$ (comme suggéré par Rousseeuw & Van Driessen, 1999). Pour obtenir une meilleure efficacité, nous pourrions également choisir $\alpha = 0.25$ ce qui donne un point de rupture de 25%.

3. Les Fonctionnelles Statistiques

Soit F la fonction de répartition de la variable Z. La distribution F sera supposée normale multivariée $N_d(\mu, \Sigma)$ où $\mu \in I\!R^d$ et $\Sigma \in$ MDPS(d), l'ensemble de toutes les matrices carrées de dimension d définies positives et symétriques. Néanmoins, les résultats obtenus seront directement généralisables au cas des distributions elliptiques.

Soit $\mathcal{F}$ l'ensemble de toutes les distributions sur $I\!R^{p+q}$ ou un large sous ensemble de celui-ci. Alors, une fonctionnelle statistique correspondant à un estimateur multivarié de dispersion est une application $C : \mathcal{F} \to MDPS(p+q)$ qui envoie une distribution arbitraire $G \in \mathcal{F}$ vers $C(G)$. Le problème d'estimation consiste à estimer $C(F)$ où F est la distribution population. Un estimateur naturel de $C(F)$ est donné par $C_n = C(F_n)$ où F_n est la distribution échantillon. Nous utiliserons la notation $C(Z)$ au lieu de $C(G)$ quand Z suit la distribution G. Par abus de langage, nous dirons parfois l'estimateur C au lieu de la fonctionnelle C. Les fonctionnelles associées aux estimateurs des valeurs propres de $\mathcal{M}_x$ et $\mathcal{M}_y$ seront notées l_j et les fonctionnelles des vecteurs propres seront appelées a_j et b_j pour $j = 1, \ldots, p$. Donc $l_j(G)$, $a_j(G)$ et $b_j(G)$ sont les valeurs et vecteurs propres de $M_x(G)$ et $M_y(G)$, où

$$\begin{aligned} M_x(G) &= C_{XX}^{-1}(G)C_{XY}(G)C_{YY}^{-1}(G)C_{YX}(G) \\ M_y(G) &= C_{YY}^{-1}(G)C_{YX}(G)C_{XX}^{-1}(G)C_{XY}(G), \end{aligned} \tag{5}$$

et

$$C(G) = \begin{bmatrix} C_{XX}(G) & C_{XY}(G) \\ C_{YX}(G) & C_{YY}(G) \end{bmatrix}$$

pour chaque $G \in \mathcal{F}$. Les fonctionnelles associées aux estimateurs des corrélations canoniques seront notées r_j, donc $r_j(G) = \sqrt{l_j(G)}$.

Nous n'étudions que des estimateurs multivariés de dispersion C convergents au sens de Fisher pour des distributions normales, ce qui revient à dire que $C(F) = \Sigma$. Nous supposerons aussi que les estimateurs multivariés de dispersion C sont affins équivariants, c'est-à-dire que $C(AX + b) = AC(X)A^t$ pour tout $b \in I\!R^d$ et pour chaque matrice carrée A de dimension d et non singulière. Il résulte que $M_x(F) = \mathcal{M}_x$, $M_y(F) = \mathcal{M}_y$ et les estimateurs a_j,b_j et r_j seront donc aussi convergents au sens de Fisher, c'est-à-dire que

$$a_j(F) = \alpha_j, \qquad b_j(F) = \beta_j, \qquad r_j(F) = \rho_j$$

où les paramètres α_j, β_j et ρ_j ont été définis à la première section. Si $q > p$, nous posons $\rho_{p+1} = \ldots = \rho_q = 0$ et $\beta_{p+1}, \ldots, \beta_q$ sont les vecteurs propres de la matrice $\mathcal{M}_y$ correspondant aux valeurs propres nulles (ces vecteurs propres sont uniques à une transformation orthogonale près).

Nous allons maintenant introduire une notion importante en robustesse, la fonction d'influence. Plus d'information sur les fonctionnelles et les fonctions d'influence sont disponibles dans Hampel *et al.* (1986).

DÉFINITION 1. — *Soit T une fonctionnelle statistique, la fonction d'influence de T au point z en la distribution F, est définie par*

$$IF(z,T,F) = \lim_{\varepsilon \downarrow 0} \frac{T((1-\varepsilon)F + \varepsilon\Delta_z) - T(F)}{\varepsilon} = \frac{\partial}{\partial \varepsilon}\left(T(F_\varepsilon)\right)|_{\varepsilon=0},$$

où Δ_z est la distribution de Dirac qui met toute sa masse au point z, et $F_\varepsilon = (1-\varepsilon)F + \varepsilon\Delta_z$.

La fonction d'influence $IF(z,T,F)$ donne une mesure de l'influence du point z sur la fonctionnelle T si la distribution sous-jacente est F. On s'attend donc pour un estimateur robuste à une fonction d'influence bornée. Pour calculer les fonctions d'influence des corrélations canoniques et vecteurs canoniques, nous aurons besoin du lemme suivant (Croux & Haesbroeck 2000), qui caractérise la fonction d'influence pour des estimateurs de dispersion affins équivariants.

LEMME 1. — *Pour toute fonctionnelle C associée à un estimateur multivarié de dispersion affin équivariant possédant une fonction d'influence, il existe deux fonctions γ_C et δ_C : $[0,\infty[\rightarrow I\!R$ tel que*

$$IF(z,C,F) = \gamma_C(d(z))(z-\mu)(z-\mu)^t - \delta_C(d(z))\Sigma \qquad (6)$$

où $d^2(z) = (z-\mu)^t\Sigma^{-1}(z-\mu)$ et $F = N_d(\mu,\Sigma)$.

TABLEAU 1

Fonctions γ_C pour les estimateurs MCD, RMCD et l'estimateur classique (Cov)

Cov	$\gamma_{\mathrm{Cov}}(t) = 1$
MCD	$\gamma_{\mathrm{MCD}}(t) = \frac{I(t \leqslant \sqrt{q_\alpha})}{F_{\chi^2_{d+4}}(q_\alpha)}$ avec $q_\alpha = \chi^2_{d,1-\alpha}$
RMCD	$\gamma_{\mathrm{RMCD}}(t) = \frac{d_2+2d_3}{d_2}\gamma_{\mathrm{MCD}}(t) + \frac{1}{d_2}I(t \leqslant \sqrt{q_\delta})$
	où $q_\delta = \chi^2_{d,1-\delta}$, $d_2 = F_{\chi^2_{d+2}}(q_\delta)$, $d_3 = -\frac{1}{2}F_{\chi^2_{d+4}}(q_\delta)$.

Les fonctions d'influence des estimateurs MCD et RMCD ont été calculées par Croux & Haesbroeck (1999). Les fonctions γ_C, pour les 2 estimateurs robustes ainsi que pour l'estimateur échantillon de la matrice de covariance, sont décrites dans le Tableau 1. Nous verrons plus loin que les fonctions δ_C jouent un rôle moins important que les fonctions γ_C. Sur la Figure 1, nous voyons que γ_{MCD} est une fonction en escalier avec une discontinuité et γ_{RMCD} avec deux discontinuités. Ces deux fonctions deviennent nulles après un certain point. Comme ces deux fonctions sont non croissantes, leur contribution à la fonction d'influence décroit si la distance entre z et μ dans la métrique imposée par Σ augmente. La fonction γ_{Cov} étant constante, elle donne donc le même poids à tous les points.

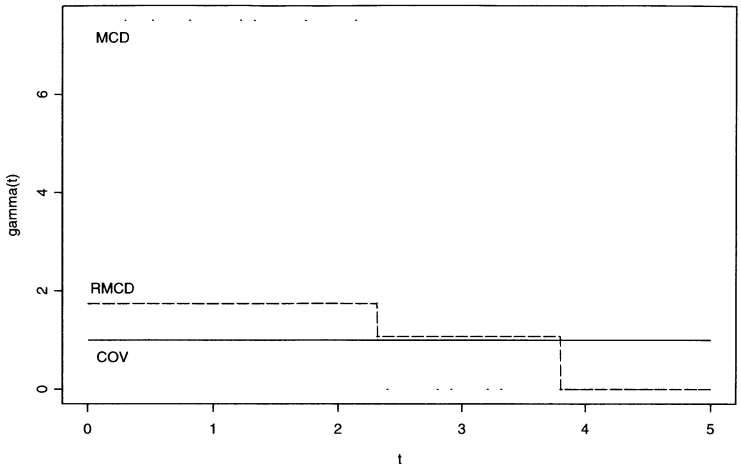

FIGURE 1

Graphiques des fonctions $\gamma(t)$ pour les estimateurs MCD (en pointillé), RMCD (en traits disjoints) et l'estimateur classique (ligne continue) pour $d = 6$, $\alpha = 0.5$ et $\delta = 0.025$

4. Fonctions d'influence des corrélations canoniques et des vecteurs canoniques

Dans la section précédente, nous avons donné une expression pour la fonction d'influence associée à des estimateurs multivariés de dispersion. A partir de ces fonctions d'influence nous allons calculer les fonctions d'influence des corrélations canoniques et des vecteurs canoniques dans un cadre tout à fait général, c'est-à-dire avec l'emploi de n'importe quel estimateur régulier de la matrice de covariance. Pour calculer les fonctions d'influence de M_x et M_y, nous avons besoin du lemme suivant.

LEMME 2. — *La fonction d'influence de M_x en la distribution F, définie par (5), est donnée par*

$$\begin{aligned} IF(z, M_x, F) =& \gamma(d(z))\Sigma_{XX}^{-1}\{(x-\mu_x)(y-\mu_y)^t\Sigma_{YY}^{-1}\Sigma_{YX} \\ &+ \Sigma_{XY}\Sigma_{YY}^{-1}(y-\mu_y)(x-\mu_x)^t - (x-\mu_x)(x-\mu_x)^t\mathcal{M}_x \\ &- \Sigma_{XY}\Sigma_{YY}^{-1}(y-\mu_y)(y-\mu_y)^t\Sigma_{YY}^{-1}\Sigma_{YX}\} \end{aligned}$$

où $\gamma(.)$ est la fonction qui apparaît dans la fonction d'influence de la fonctionnelle C. La fonction d'influence de M_y est donnée de manière similaire.

Les fonctions d'influence des corrélations canoniques et des vecteurs canoniques découleront des fonctions d'influence de M_x et M_y et ceci grâce à une variante d'un lemme de Sibson (1979). Ce lemme nous donne les fonctions d'influence des valeurs et vecteurs propres de toute matrice symétrique définie positive, sur base de la

fonction d'influence de la matrice elle-même. Puisque les matrices M_x et M_y ne sont pas symétriques, nous devons donc adapter le lemme de Sibson, en tenant également compte des contraintes de normalisation (2). Nous supposerons que toutes les valeurs propres sont distinctes et strictement positives.

LEMME 3. — *Soit (λ_j, α_j) les paires de valeurs propres et vecteurs propres d'une matrice $\mathcal{M}$ quelconque telles que $\alpha_i^t \Sigma \alpha_j = \delta_{ij}$ pour $i, j = 1, \ldots, p$ et Σ une matrice définie positive quelconque. Alors*

$$IF(z, l_j, F) = \alpha_j^t \Sigma IF(z, M, F) \alpha_j \tag{7}$$

$$IF(z, a_j, F) = \sum_{\substack{k=1 \\ k \neq j}}^{p} \left(\frac{\alpha_k^t \Sigma IF(z, M, F) \alpha_j}{\lambda_j - \lambda_k} \right) \alpha_k - \frac{1}{2} \left(\alpha_j^t IF(z, C, F) \alpha_j \right) \alpha_j. \tag{8}$$

où M est une fonctionnelle tel que $M(F) = \mathcal{M}$, l_j et a_j sont les valeurs et vecteurs propres de M, avec $a_j^t C a_j = 1$, C étant une fonctionnelle telle que $C(F) = \Sigma$.

Le théorème suivant va nous fournir les fonctions d'influence des corrélations canoniques et des vecteurs canoniques dans un cadre très général. Pour plus de lisibilité, nous allons adopter les notations suivantes,

$$u_k = \alpha_k^t (x - \mu_x), \qquad v_k = \beta_k^t (y - \mu_y)$$

pour $k = 1, \ldots, p$. Les vecteurs u_k et v_k sont respectivement appelés scores de x sur le k-ième vecteur canonique α_k et scores de y sur le k-ième vecteur canonique β_k.

THÉORÈME 1. — *Les fonctions d'influence des estimateurs des corrélations canoniques et des vecteurs canoniques basés sur un estimateur C affin équivariant de la matrice de covariance en la distribution $F = N_d(\mu, \Sigma)$, sont données par*

$$IF(z, r_j, F) = \gamma(d(z))(u_j v_j - \frac{1}{2}\rho_j u_j^2 - \frac{1}{2}\rho_j v_j^2), \tag{9}$$

$$IF(z, a_j, F) = \gamma(d(z)) \sum_{\substack{k=1 \\ k \neq j}}^{p} \{ \frac{\rho_j (v_j - \rho_j u_j) u_k + \rho_k (u_j - \rho_j v_j) v_k}{\rho_j^2 - \rho_k^2} \} \alpha_k \tag{10}$$

$$+ \frac{1}{2} \left(\delta(d(z)) - \gamma(d(z)) u_j^2 \right) \alpha_j$$

et

$$IF(z, b_j, F) = \gamma(d(z)) \sum_{\substack{k=1 \\ k \neq j}}^{q} \{ \frac{\rho_j (u_j - \rho_j v_j) v_k + \rho_k (v_j - \rho_j u_j) u_k}{\rho_j^2 - \rho_k^2} \} \beta_k \tag{11}$$

$$+ \frac{1}{2} \left(\delta(d(z)) - \gamma(d(z)) v_j^2 \right) \beta_j$$

où γ et δ sont les fonctions qui apparaissent en (6), $d^2(z) = (z-\mu)^t\Sigma^{-1}(z-\mu)$ et $\Sigma = C(F)$.

Pour aider à visualiser les fonctions d'influence et le caractère borné de celles-ci dans le cas robuste, nous allons les représenter graphiquement. Malheureusement, nous devons nous restreindre à un graphique en 3 dimensions maximum. Nous avons donc calculé la fonction d'influence de la première corrélation canonique r_1 dans le cas particulier où $p = q = 1$. Nous avons envisagé 3 procédures d'estimation : soit de façon classique, soit par la méthode MCD ou soit par la méthode RMCD. Les dessins de la Figure 2 donnent sur l'axe vertical la valeur de la fonction d'influence au point $z = (x, y)$. Il est maintenant très facile de constater que dans le cas classique, la fonction d'influence de r_1 en la distribution $F = \mathrm{N}_2(0, \Sigma)$ où $\Sigma = ((1, 0.5)^t, (0.5, 1)^t)$ n'est pas bornée, ce qui signifie qu'un point aberrant peut avoir une influence infinie sur l'estimateur. Par contre dans les cas MCD et RMCD, les fonctions d'influence sont bornées. Donc, si un point est vraiment trop loin de la grande majorité des données, il n'a plus d'influence sur l'estimateur r_1. Par contre si le point n'est pas trop loin, nous en tiendrons compte et son influence sera contrôlée. Nous voyons que la fonction d'influence des estimateurs basés sur la méthode RMCD a un niveau de cassure supplémentaire par rapport à celle basée sur la méthode MCD. En ce qui concerne les vecteurs canoniques, il est aisé de voir à partir du Théorème 1 que a_j et b_j seront bornés si la fonction $\gamma(x)$ s'annule pour tout x supérieur à un certain x_0 et si la fonction δ est bornée.

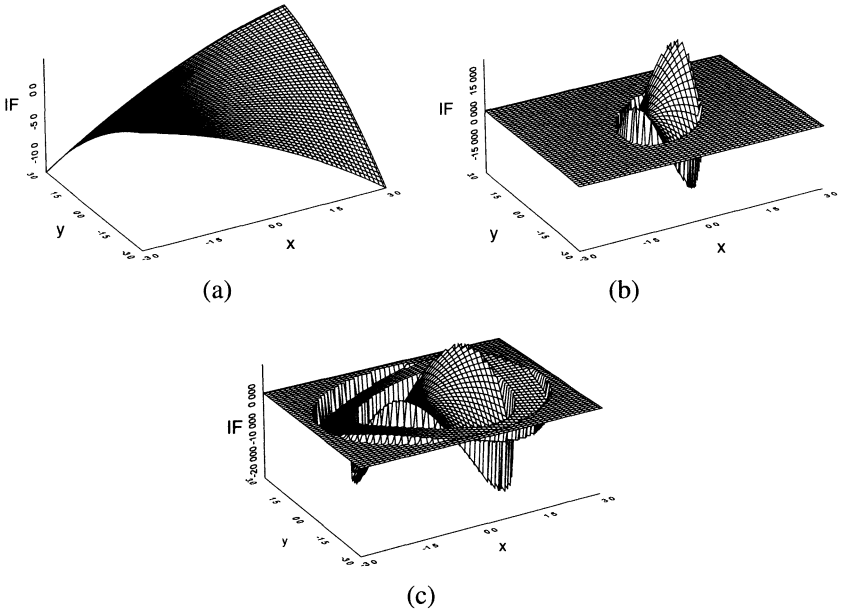

FIGURE 2

Fonction d'influence pour la première corrélation canonique calculée sur base de (a) l'estimateur classique (b) l'estimateur MCD et (c) l'estimateur RMCD

Dans le cas particulier où $\gamma \equiv 1$ et $\delta \equiv 1$, nous obtenons les fonctions d'influence des corrélations canoniques et vecteurs canoniques où l'estimateur de la matrice covariance n'est rien d'autre que l'estimateur de covariance empirique. Ces fonctions d'influence sont non bornées ce qui indique que ces estimateurs sont non robustes. On peut vérifier que cela correspond aux résultats obtenus par Romanazzi (1992), qui a obtenu les fonctions d'influence dans le cas classique. Les liens entre les fonctions d'influence des corrélations canoniques et des vecteurs canoniques utilisant un estimateur robuste et l'estimateur empirique classique, sont extrêmement simples et sont donnés par le corollaire suivant, qui découle immédiatement du Théorème 1.

COROLLAIRE 1. — *Les liens entre les fonctions d'influence des corrélations canoniques et des vecteurs canoniques utilisant un estimateur robuste et l'estimateur empirique classique (*Cov*) sont les suivants :*

$$IF(z, l_j, F) = \gamma(d(z))IF(z, l_j^{\mathrm{Cov}}, F)$$
$$IF(z, a_j, F) = \gamma(d(z))IF(z, a_j^{\mathrm{Cov}}, F) + \frac{1}{2}(\delta(d(z)) - \gamma(d(z)))\alpha_j$$
$$IF(z, b_j, F) = \gamma(d(z))IF(z, b_j^{\mathrm{Cov}}, F) + \frac{1}{2}(\delta(d(z)) - \gamma(d(z)))\beta_j.$$

pour $j = 1, \ldots, p$, où γ et δ sont les fonctions qui apparaîssent en (6).

En pratique, on préfère parfois faire l'analyse canonique basée sur la matrice des corrélations

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

au lieu de $C(G)$. La matrice de corrélation robuste est définie à partir de $C(G)$ comme dans le cas classique. Nous avons par exemple que

$$R_{XX}(G) = D_X^{-1/2}(G) C_{XX}(G) D_X^{-1/2}(G),$$

où $D_X(G) = \mathrm{diag}(C_{XX}(G))$ reprend seulement les éléments diagonaux de $C_{XX}(G)$ et annule les autres. En utilisant R à la place de C, les corrélations canoniques restent identiques mais les vecteurs canoniques changent. Néanmoins, il existe un lien simple entre les deux versions. Désignons $\tilde{a}_j(G)$ et $\tilde{b}_j(G)$ les fonctionnelles pour les vecteurs canoniques basés sur la matrice de corrélation. Alors

$$\tilde{a}_j(G) = D_X^{1/2}(G) a_j(G) \tag{12}$$
$$\tilde{b}_j(G) = D_Y^{1/2}(G) b_j(G)$$

pour chaque $G \in \mathcal{F}$, et $1 \leqslant j \leqslant p$. A partir de (12) et du Théorème 1, nous retrouvons les courbes d'influence des estimateurs des corrélations canoniques et des vecteurs canoniques basés sur un estimateur R de la matrice de corrélation.

THÉORÈME 2. — *Soit R un estimateur de la matrice de corrélation déduit d'un estimateur C affin équivariant de la matrice de covariance. La fonction d'influence*

des estimateurs des vecteurs canoniques basés sur R en la distribution $F = N_d(\mu, \Sigma)$, est donnée par

$$IF(z, \tilde{a}_j, F) = \gamma(d(z)) \left(\sum_{\substack{k=1 \\ k \neq j}}^{p} \{ \frac{\rho_j(v_j - \rho_j u_j)u_k + \rho_k(u_j - \rho_j v_j)v_k}{\rho_j^2 - \rho_k^2} \} \tilde{\alpha}_k + \frac{1}{2}(D_{\tilde{x}} \tilde{\alpha}_j - u_j^2 \tilde{\alpha}_j) \right) \quad (13)$$

où γ est la fonction qui apparaît en (6), $\tilde{\alpha}_k$ est le vecteur canonique au niveau de la population (donc $\tilde{a}_j(F) = \tilde{\alpha}_j$) et $D_{\tilde{x}} = \mathrm{diag}(\tilde{x}\tilde{x}^t)$ avec $\tilde{x} = \mathrm{diag}(\Sigma_{XX})^{-1/2}(x - \mu_x)$. Bien évidemment, une expression analogue est valide pour $IF(z, \tilde{b}_j, F)$.

Les fonctions d'influence mesurent la robustesse d'un estimateur, mais elles peuvent aussi être utilisées pour obtenir une expression pour la matrice de variance-covariance des estimateurs. Si l'estimateur de la matrice de covariance est asymptotiquement normal (ce qui est le cas pour les estimateurs MCD et RMCD), les estimateurs dérivés des corrélations et vecteurs canoniques le seront également. En effet, les vecteurs et valeurs propres des matrices M_x et M_y sont des fonctions différentiables des éléments de la matrice C. Par la méthode Delta, on obtiendra alors

$$\sqrt{n}(\hat{\alpha}_j - \alpha_j) \xrightarrow{d} N_d(0, ASV(\hat{\alpha}_j))$$

où $ASV(\hat{\alpha}_j) = E_F[IF(z, a_j, F)IF(z, a_j, F)^t]$ et $\hat{\alpha}_j = a_j(F_n)$ avec F_n la distribution empirique pour $1 \leqslant j \leqslant p$. Un estimateur de la matrice de covariance de l'estimateur $\hat{\alpha}_j$ est alors

$$\mathrm{Cov}(\hat{\alpha}_j) \approx \frac{1}{n^2} \sum_{i=1}^{n} IF(z_i, a_j, \hat{F}_n) IF(z_i, a_j, \hat{F}_n)^t \quad (14)$$

avec $\hat{F}_n$ la fonction de répartition de la loi normale $N(T_n, C_n)$, où T_n et C_n sont les estimateurs de localisation et de dispersion multivariées. Des expressions similaires sont valides pour $\mathrm{Cov}(\hat{\beta}_j)$, $\mathrm{Cov}(\hat{\tilde{\alpha}}_j)$, $\mathrm{Cov}(\hat{\tilde{\beta}}_j)$ et $Var(\hat{\rho}_j)$, avec $\hat{\beta}_j = b_j(F_n)$, $\hat{\tilde{\alpha}}_j = \tilde{a}_j(F_n)$, $\hat{\tilde{\beta}}_j = \tilde{b}_j(F_n)$ et $\hat{\rho}_j = r_j(F_n)$. Ainsi

$$Var(\hat{\rho}_j) \approx \frac{1}{n^2} \sum_{i=1}^{n} IF(z_i, r_j, \hat{F}_n)^2 \quad (15)$$

pour $1 \leqslant j \leqslant p$.

La distribution asymptotique des corrélations et vecteurs canonique constitue le sujet de nombreux papiers (*cf.* les références citées dans Anderson (1999)). Le premier papier sur ce sujet est celui de Hsu (1941) qui donne des résultats pour les corrélations canoniques. Un papier plus récent est celui de Anderson cité ci-dessus qui donne la distribution asymptotique pour les vecteurs canoniques. Tous ces résultats sont restreints à l'analyse canonique basée sur la matrice variance-covariance échantillon

et nombre d'entre eux exigent la normalité. Le papier de Eaton & Tyler (1994) est une des exceptions qui donnent des résultats pour des distributions elliptiques. Nous rappellons que le résultat que nous avons déduit est valable pour une analyse canonique basée sur n'importe quel estimateur affin et convergent de la matrice de dispersion Σ. En plus, nous avons donné des résultats pour l'analyse canonique basée sur les matrices de corrélation, cas qui ne semble pas avoir été très étudié dans la littérature.

5. Exemple

Nous allons illustrer cette théorie par l'étude de la base de données sur la condition physique de Linnerud (traitée dans Tenenhaus 1998, page 15). L'étude porte sur les liens entre 3 variables d'exercices (X_1 = Poids, X_2 = Tour de taille, X_3 = Pouls) et 3 variables de mesures physiques (Y_1 = Nombre de tractions à la barre fixe, Y_2 = Nombre de flexions, Y_3 = Nombre de sauts) pour 20 individus. La première étape de cette étude empirique sera consacrée à la détection des points aberrants. Ensuite, nous calculerons et comparerons les résultats obtenus par l'analyse canonique classique et robuste. Nous utiliserons comme estimateurs robustes les estimateurs RMCD avec un point de rupture de 50%, et un paramètre δ égal à 0.025.

Pour détecter les éventuels points aberrants dans un échantillon $Z = \{z_1, \ldots, z_n\} \in I\!R^d$, nous utilisons la distance robuste

$$d_i = \sqrt{(z_i - T_n(Z))^t C_n(Z)^{-1}(z_i - T_n(Z))}$$

où T_n et C_n sont les estimateurs multivariés robustes. Si nous prenons les estimateurs classiques, nous obtenons la distance usuelle de Mahalanobis. Sur la Figure 3, nous avons repris les distances de Mahalanobis et les distances robustes. L'observation 10 apparaît très clairement comme un point aberrant, ce qui est beaucoup moins visible pour les distances de Mahalanobis. En effet, les distances de Mahalanobis classiques souffrent du «masking effect» (Rousseeuw & van Zomeren, 1990).

Nous allons estimer les corrélations canoniques et les vecteurs canoniques au moyen de l'estimation de la matrice de corrélation par la méthode classique et par la méthode robuste. Le Tableau 2 fournit les matrices des corrélations calculées à partir des estimations classiques (au dessus de la diagonale) et RMCD (en dessous de la diagonale). Il est frappant de voir dans le Tableau 2 que 3 corrélations ont des signes différents suivant la méthode d'estimation. Par la méthode classique, les corrélations entre d'une part le pouls et d'autre part le nombre de tractions, de flexions ou de sauts sont positives, alors que dans le cas robuste ces 3 corrélations sont négatives. Or il est logique d'affirmer qu'une personne au rythme cardiaque plus lent aura plus de facilité pour faire des exercices physiques, ce qui accrédite la méthode robuste.

Les estimations des corrélations canoniques et vecteurs canoniques basées sur la matrice de corrélation se trouvent respectivement dans les Tableaux 3 et 4 ainsi que leurs écarts types. Dans le Tableau 3, les écarts types sont calculés en utilisant (15), et pour le Tableau 4, on a employé l'expression (14) pour $\hat{\tilde{\alpha}}_j$.

Nous remarquons que les deux premières corrélations canoniques robustes sont plus grandes que celles obtenues par la méthode classique. La plus grande différence

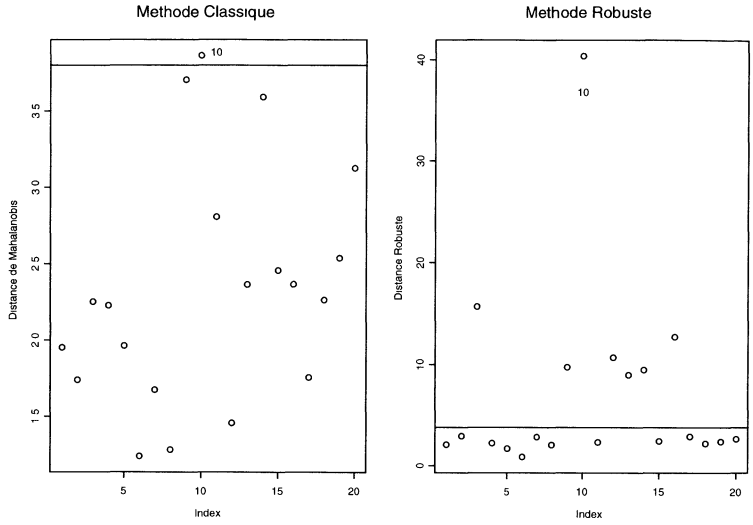

FIGURE 3

Distances de Mahalanobis (graphique de gauche) et distances robustes (graphique de droite) pour les données de Linnerud

TABLEAU 2

Matrice des corrélations obtenue d'une façon classique (au dessus de la diagonale) et robuste (en dessous de la diagonale)

	Poids	Tour de taille	Pouls	Tractions	Flexions	Sauts
Poids	1	0.870	-0.366	-0.390	-0.493	-0.226
Tour de taille	0.844	1	-0.353	-0.552	-0.646	-0.191
Pouls	-0.591	-0.239	1	0.151	0.225	0.035
Tractions	-0.157	-0.435	-0.304	1	0.696	0.496
Flexions	-0.347	-0.675	-0.151	0.503	1	0.669
Sauts	-0.212	-0.505	-0.223	0.276	0.926	1

se reflète dans $\hat{\rho}_2$ qui passe de 0.201 à 0.559. En ce qui concerne les vecteurs canoniques (Tableau 4), nous avons pris comme convention pour les signes que la première coordonnée de chaque vecteur canonique, soit positive. De grandes différences entre les valeurs des vecteurs propres apparaissent et même les signes varient suivant la méthode utilisée. De plus, les écarts types sont relativement grands (dans les cas classique et robuste) et donc peu de coefficients sont significativement différents de zéro. Une partie des interprétations seront donc soumises à caution. Ce fait est

TABLEAU 3

Corrélations canoniques estimées par la méthode classique et la méthode robuste (basée sur RMCD) pour les données de Linnerud. Les écarts types sont donnés entre parenthèses

	Classique		RMCD	
ρ_1	0.796	(0.066)	0.836	(0.057)
ρ_2	0.201	(0.186)	0.559	(0.436)
ρ_3	0.073	(0.210)	0.038	(0.129)

TABLEAU 4

Vecteurs canoniques estimés par la méthode classique et la méthode robuste (basée sur RMCD) pour les données de Linnerud. Les écarts types pour chaque composante des vecteurs estimés sont donnés entre parenthèses

	Classique			RMCD		
$\tilde{\alpha}_1$	0.775	-1.579	0.059	0.944	-1.666	0.046
	(0.367)	(0.291)	(0.168)	(0.421)	(0.299)	(0.397)
$\tilde{\alpha}_2$	1.884	-1.181	0.231	0.573	-0.474	1.203
	(0.657)	(0.948)	(0.907)	(1.105)	(0.798)	(0.437)
$\tilde{\alpha}_3$	0.191	-0.506	-1.051	2.583	-1.563	0.978
	(1.898)	(1.378)	(0.193)	(0.454)	(0.479)	(0.400)
$\tilde{\beta}_1$	0.349	1.054	-0.716	0.070	1.615	-0.760
	(0.084)	(0.146)	(0.166)	(0.426)	(1.009)	(0.911)
$\tilde{\beta}_2$	0.376	-0.123	-1.062	1.264	-3.216	2.900
	(1.339)	(1.307)	(0.455)	(0.267)	(1.314)	(1.132)
$\tilde{\beta}_3$	1.297	-1.237	0.419	0.659	0.588	-1.325
	(0.396)	(0.352)	(1.023)	(0.433)	(1.174)	(1.006)

expliqué par le faible nombre d'observations. L'analyse présentée dans cet exemple est donc plutôt exploratoire qu'inférentielle.

Nous allons être très succints pour l'interprétation des résultats. Notre but est de déterminer la contribution de chaque variable dans les composantes canoniques. Deux écoles de pensée s'opposent dans ce domaine. Rencher (1998) est partisan d'utiliser les coordoonnées des vecteurs canoniques qui mesurent l'apport marginal de chaque

variable à la construction des composantes canoniques de leur groupe et c'est ce qui est nécessaire pour prendre en compte l'aspect multivarié de l'analyse. Pour notre exemple, les premières variables canoniques robustes estimées s'écrivent

$$U_1 = 0.944\tilde{X}_1 - 1.666\tilde{X}_2 + 0.046\tilde{X}_3$$

où $\tilde{X}_1$, $\tilde{X}_2$ et $\tilde{X}_3$ sont respectivement les variables centrées réduites (par la méthode RMCD) du poids, du tour de taille et du pouls

$$V_1 = 0.070\tilde{Y}_1 + 1.615\tilde{Y}_2 - 0.760\tilde{Y}_3$$

où $\tilde{Y}_1$, $\tilde{Y}_2$ et $\tilde{Y}_3$ sont respectivement les variables centrées réduites du nombre de tractions à la barre fixe, du nombre de flexions et du nombre de sauts. Nous remarquons que la variable Pouls ne contribue pas beaucoup à la construction de la première variable canonique U_1. L'interprétation est difficile car les vecteurs canoniques opposent des variables positivement corrélées entre elles (par exemple poids et taille). Tenenhaus (page 18, 1998) propose alors d'utiliser les corrélations entre les variables et les variables canoniques. Les corrélations robustes entre les mesures physiques (X_1, X_2 et X_3) et leurs variables canoniques (U_1, U_2 et U_3) apparaissent dans le Tableau 5. De même, on trouvera les corrélations entre les exercices (Y_1, Y_2 et Y_3) et leurs variables canoniques (V_1, V_2 et V_3). Nous pouvons penser que U_1 est une mesure de mauvaise santé, tandis que V_1 oppose la force physique à l'endurance (basée sur la méthode robuste). Nous remarquons de très grandes différences entre les méthodes classique et robuste au niveau des corrélations entre V_1 et les variables du deuxième groupe. Nous n'allons pas analyser les variables canoniques d'ordre supérieur à deux dans cet exemple car l'interprétation serait difficile.

Pour clôturer cette interprétation, la Figure 4 permet de visualiser la liaison entre les composantes canoniques U_1 et V_1 calculées à partir des méthodes classique et robuste. La dispersion des points autour de la droite ajustée est plus grande dans le cas classique que dans le cas robuste. La corrélation robuste entre U_1 et V_1 est positive et assez grande (0.851), et nous voyons que l'individu 10 est décentré. Le graphique associé à la méthode classique ne permet pas de détecter le comportement atypique de l'individu 10. Néanmoins rappelons que le but de ce graphique n'est pas de découvrir les points aberrants, ce rôle étant rempli par la détermination des distances robustes. Il semble que les variables canoniques estimées d'une façon classique ont été attirées par le point 10. Il est donc préférable de baser l'interprétation sur l'analyse robuste, qui nous protège contre le dixième individu qui est un point aberrant.

6. Conclusion

Le traitement des points aberrants dans l'analyse canonique a déjà été étudié dans la littérature, par Campbell (1982) et Karnel (1991), en utilisant les M-estimateurs multivariés. Nous avons étendu cette étude en utilisant les estimateurs multivariés à haut point de rupture et nous donnons une expression explicite des fonctions d'influence des corrélations canoniques et vecteurs canoniques basés sur n'importe quel estimateur convergent et affin équivariant de la matrice de variance-covariance.

TABLEAU 5

Corrélations entre le premier groupe de variables et leurs composantes canoniques ainsi que pour le deuxième groupe de variables et leurs composantes canoniques pour les méthodes classique et robuste

Méthode Classique :	U_1	U_2	U_3		V_1	V_2	V_3
Poids	-0.620	0.772	0.135	Tractions	0.728	-0.237	0.644
Tour de taille	-0.925	0.378	0.031	Flexions	0.818	-0.573	-0.054
Pouls	0.333	-0.041	-0.942	Sauts	0.162	-0.959	0.234
Méthode robuste :							
Poids	-0.694	-0.352	0.226	Tractions	0.146	0.415	-0.246
Tour de taille	-0.959	-0.477	-0.193	Flexions	0.236	0.411	-0.452
Pouls	0.327	0.960	0.502	Sauts	-0.564	0.946	-0.956

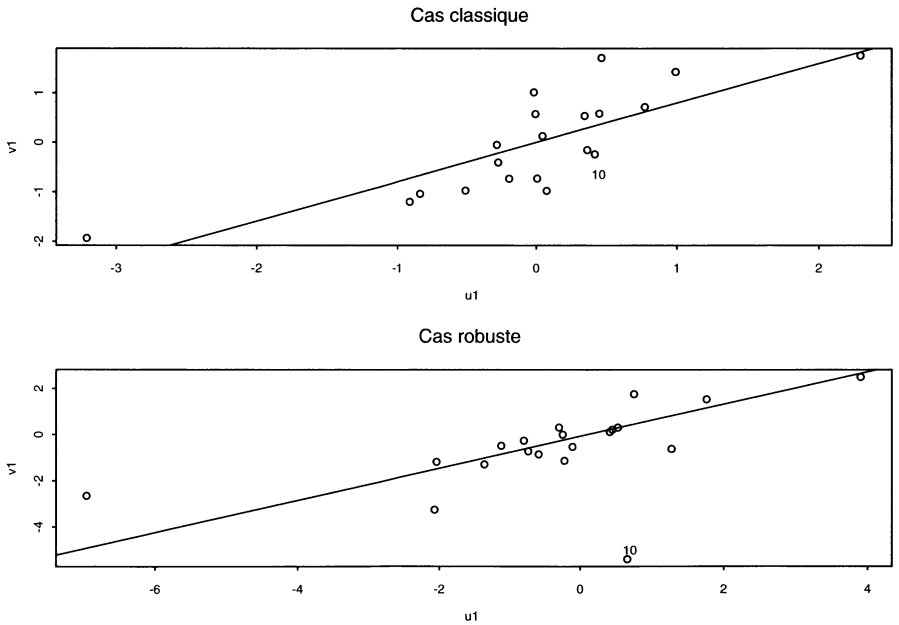

FIGURE 4

Graphiques des variables canoniques U_1 et V_1 estimées par la méthode classique et robuste (basée sur RMCD) pour les données de Linnerud

7. Annexes

Démonstration du Lemme 2 :

Pour démontrer ce lemme, nous utilisons le résultat suivant (Pullman 1976, page 120) :

$$IF(z, C^{-1}, F) = -\Sigma^{-1} IF(z, C, F)\Sigma^{-1}. \tag{16}$$

Par dérivation et en utilisant le résultat ci-dessus avec $F_\varepsilon = (1-\varepsilon)F + \varepsilon\Delta_z$, nous trouvons :

$$\begin{aligned} IF(z, M_x, F) &= \frac{\partial}{\partial\varepsilon} M_x(F_\varepsilon) \mid_{\varepsilon=0} \\ &= \Sigma_{XX}^{-1}\{IF(z, C_{XY}, F) - IF(z, C_{XX}, F)\Sigma_{XX}^{-1}\Sigma_{XY}\}\Sigma_{YY}^{-1}\Sigma_{YX} \\ &+ \Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1}\{IF(z, C_{YX}, F) - IF(z, C_{YY}, F)\Sigma_{YY}^{-1}\Sigma_{YX}\}. \end{aligned}$$

En réutilisant l'expression générale pour la fonction d'influence pour un estimateur de la matrice covariance (6) et les équations (3) et (4), le lemme est demontré.

Démonstration du Lemme 3 :

Puisque $M(G)a_j(G) = l_j(G)a_j(G)$ pour toute distribution G, nous avons pour $F_\varepsilon = (1-\varepsilon)F + \varepsilon\Delta_z$ où Δ_z est une distribution mettant toute sa masse au point $z = (x, y)$:

$$\begin{aligned} &\frac{\partial}{\partial\varepsilon} M(F_\varepsilon) \mid_{\varepsilon=0} a_j(F) + M(F)\frac{\partial}{\partial\varepsilon} a_j(F_\varepsilon) \mid_{\varepsilon=0} \\ &\qquad = \frac{\partial}{\partial\varepsilon} l_j(F_\varepsilon) \mid_{\varepsilon=0} a_j(F) + l_j(F)\frac{\partial}{\partial\varepsilon} a_j(F_\varepsilon) \mid_{\varepsilon=0}. \end{aligned}$$

Il en découle

$$IF(z, M, F)\alpha_j + \mathcal{M}\, IF(z, a_j, F) = IF(z, l_j, F)\alpha_j + \lambda_j IF(z, a_j, F),$$

et donc

$$(\mathcal{M}\Sigma^{-1} - \lambda_j\Sigma^{-1})\Sigma IF(z, a_j, F) = IF(z, l_j, F)\alpha_j - \Sigma^{-1}\Sigma IF(z, M, F)\alpha_j.$$

Il est facile de voir que $\mathcal{M}\Sigma^{-1} = (\sum_{k=1}^{p} \lambda_k\alpha_k\alpha_k^t)$ et

$$\Sigma^{-1} = \sum_{k=1}^{p} \alpha_k\alpha_k^t. \tag{17}$$

Grâce à ces deux observations, nous avons

$$(\sum_{k=1}^{p} \lambda_k \alpha_k \alpha_k^t - \lambda_j \sum_{k=1}^{p} \alpha_k \alpha_k^t) \Sigma IF(z, a_j, F)$$

$$= IF(z, l_j, F)\alpha_j - \sum_{k=1}^{p} \alpha_k \alpha_k^t \Sigma IF(z, M, F)\alpha_j,$$

et donc

$$\sum_{\substack{k=1 \\ k \neq j}}^{p} \{(\lambda_k - \lambda_j)\alpha_k^t \Sigma IF(z, a_j, F) + \alpha_k^t \Sigma IF(z, M, F)\alpha_j\}\alpha_k$$

$$-\{IF(z, l_j, F) - \alpha_j^t \Sigma IF(z, M, F)\alpha_j\}\alpha_j = 0.$$

Comme les vecteurs propres α_k sont linéairement indépendants (mais non orthogonaux), les coefficients devant les vecteurs propres dans l'équation ci-dessus doivent être nuls ce qui implique déjà l'équation (7) et

$$\alpha_k^t \Sigma IF(z, a_j, F) = \frac{\alpha_k^t \Sigma IF(z, M, F)\alpha_j}{\lambda_j - \lambda_k}$$

pour tout $k \neq j$. Mais, grâce à (17), nous pouvons écrire tout point x comme suit $x = \sum_{k=1}^{p} (x^t \Sigma \alpha_k)\alpha_k$. Donc, nous obtenons

$$IF(z, a_j, F) = \sum_{\substack{k=1 \\ k \neq j}}^{p} \left(\frac{\alpha_k^t \Sigma IF(z, M, F)\alpha_j}{\lambda_j - \lambda_k} \right) \alpha_k + \alpha_j^t \Sigma IF(z, a_j, F)\alpha_j. \quad (18)$$

De plus, la restriction $a_j^t C(F_\varepsilon) a_j = 1$ implique

$$2IF(z, a_j, F)^t \Sigma \alpha_j = -\alpha_j^t IF(z, C, F)\alpha_j. \quad (19)$$

Ainsi, le deuxième résultat du lemme est démontré en combinant (18) et (19). □

Démonstration du Théorème 1 :

Puisque les r_j^2 sont les valeurs propres de la fonctionnelle M_x, le premier résultat du lemme 3 donne

$$IF(z, r_j^2, F) = \alpha_j^t \Sigma_{XX} IF(z, M_x, F)\alpha_j.$$

Et, comme dans le lemme 2 nous avons explicitement calculé la fonction d'influence de M_x, il vient que

$$\begin{aligned} IF(z, r_j^2, F) =& \gamma(d(z))\alpha_j^t \{(x - \mu_x)(y - \mu_y)^t \Sigma_{YY}^{-1} \Sigma_{YX} \\ &+ \Sigma_{XY} \Sigma_{YY}^{-1} (y - \mu_y)(x - \mu_x)^t - (x - \mu_x)(x - \mu_x)^t \mathcal{M}_x \\ &- \Sigma_{XY} \Sigma_{YY}^{-1} (y - \mu_y)(y - \mu_y)^t \Sigma_{YY}^{-1} \Sigma_{YX}\} \alpha_j. \end{aligned}$$

Rappelons que (ρ_j^2, α_j) sont les valeurs et vecteurs propres de $\mathcal{M}_x$ et que (ρ_j^2, β_j) sont les valeurs et vecteurs propres de $\mathcal{M}_y$, dès lors nous pouvons utiliser les deux relations

$$\mathcal{M}_x \alpha_j = \rho_j^2 \alpha_j \qquad \text{et} \qquad \Sigma_{YY}^{-1} \Sigma_{YX} \alpha_j = \rho_j \beta_j. \tag{20}$$

Donc

$$\begin{aligned} IF(z, r_j^2, F) =& \gamma(d(z)) \rho_j \{\alpha_j^t (x-\mu_x)(y-\mu_y)^t \beta_j + \beta_j^t (y-\mu_y)(x-\mu_x)^t \alpha_j \\ & - \rho_j \alpha_j^t (x-\mu_x)(x-\mu_x)^t \alpha_j - \rho_j \beta_j^t (y-\mu_y)(y-\mu_y)^t \beta_j\}. \end{aligned}$$

En utilisant la relation suivante

$$IF(z, r_j^2, F) = 2\rho_j IF(z, r_j, F)$$

et les notations adoptées, nous obtenons le premier résultat du théorème.

Le deuxième résultat du théorème s'obtient à partir du deuxième résultat du lemme 3, qui est

$$IF(z, a_j, F) = \sum_{\substack{k=1 \\ k \neq j}}^{p} \left(\frac{\alpha_k^t \Sigma_{XX} IF(z, M_x, F) \alpha_j}{\rho_j^2 - \rho_k^2} \right) \alpha_k - \frac{1}{2} (\alpha_j^t IF(z, C_{XX}, F) \alpha_j) \alpha_j.$$

Donc, l'application du lemme 2 et (6) donnent

$$\begin{aligned} IF(z, a_j, F) = \sum_{\substack{k=1 \\ k \neq j}}^{p} & \frac{\gamma(d(z))}{\rho_j^2 - \rho_k^2} \{\alpha_k^t (x-\mu_x)(y-\mu_y)^t \Sigma_{YY}^{-1} \Sigma_{YX} \alpha_j \\ & + \alpha_k^t \Sigma_{XY} \Sigma_{YY}^{-1} (y-\mu_y)(x-\mu_x)^t \alpha_j - \alpha_k^t (x-\mu_x)(x-\mu_x)^t \mathcal{M}_x \alpha_j \\ & - \alpha_k^t \Sigma_{XY} \Sigma_{YY}^{-1} (y-\mu_y)(y-\mu_y)^t \Sigma_{YY}^{-1} \Sigma_{YX} \alpha_j\} \alpha_k \\ & - \frac{1}{2} [\alpha_j^t \{\gamma(d(z))(x-\mu_x)(x-\mu_x)^t - \delta(d(z)) \Sigma_{XX}\} \alpha_j] \alpha_j. \end{aligned}$$

En utilisant les équations (2) et (20), nous obtenons

$$\begin{aligned} IF(z, a_j, F) =& \gamma(d(z)) \sum_{\substack{k=1 \\ k \neq j}}^{p} \frac{1}{\rho_j^2 - \rho_k^2} \{\rho_j \alpha_k^t (x-\mu_x)(y-\mu_y)^t \beta_j \\ & + \rho_k \beta_k^t (y-\mu_y)(x-\mu_x)^t \alpha_j \\ & - \rho_j^2 \alpha_k^t (x-\mu_x)(x-\mu_x)^t \alpha_j - \rho_k \rho_j \beta_k^t (y-\mu_y)(y-\mu_y)^t \beta_j\} \alpha_k \\ & + \frac{1}{2} \{\delta(d(z)) - \gamma(d(z)) \alpha_j^t (x-\mu_x)(x-\mu_x)^t \alpha_j\} \alpha_j. \end{aligned}$$

Par définition des scores u_k et v_k, le deuxième résultat (10) est obtenu. □

Démonstration du Théorème 2 :

En utilisant l'équation (12), nous obtenons

$$IF(z, \tilde{a}_j, F) = IF(z, D_X^{1/2}, F)\alpha_j + \mathcal{D}_X^{1/2} IF(z, a_j, F), \tag{21}$$

où $\mathcal{D}_X = diag(\Sigma_{XX})$. En utilisant le lemme 1 pour la matrice diagonale $D_X^{1/2}$ et par définition de $D_{\tilde{x}}$, il suit que

$$\begin{aligned} IF(z, D_X^{1/2}, F) &= \frac{1}{2} IF(z, D_X, F)\mathcal{D}_X^{-1/2} \\ &= \frac{1}{2}\{\gamma(d(z))\mathrm{diag}((x-\mu)(x-\mu)^t) - \delta(d(z))\mathrm{diag}(\Sigma_{XX})\}\mathcal{D}_X^{-1/2} \\ &= \frac{1}{2}\{\gamma(d(z))\mathcal{D}_X^{1/2} D_{\tilde{x}} - \delta(d(z))\mathcal{D}_X^{1/2}\} \\ &= \frac{1}{2}\{\gamma(d(z))D_{\tilde{x}} - \delta(d(z))\}\mathcal{D}_X^{1/2}. \end{aligned} \tag{22}$$

Le résultat du théorème 2 s'obtient en insérant (22) et l'équation (10) du théorème 1 dans (21)

$$\begin{aligned} IF(z, \tilde{a}_j, F) =& \frac{1}{2}\{\gamma(d(z))D_{\tilde{x}} - \delta(d(z))\}\tilde{\alpha}_j \\ &+\gamma(d(z)) \sum_{\substack{k=1 \\ k\neq j}}^{p} \{\frac{\rho_j(v_j - \rho_j u_j)u_k + \rho_k(u_j - \rho_j v_j)v_k}{\rho_j^2 - \rho_k^2}\}\tilde{\alpha}_k \\ &+\frac{1}{2}\left(\delta(d(z)) - \gamma(d(z))u_j^2\right)\tilde{\alpha}_j, \end{aligned}$$

ce qui donne (13). □

8. Références

T. W. ANDERSON (1999), Asymptotic Theory for Canonical Correlation Analysis. *Journal of Multivariate Analysis* , 70, 1-29.

R. W. BUTLER, P. L. DAVIES & M. JHUN (1993), Asymptotics for the Minimum Covariance Determinant Estimator. *The Annals of Statistics,* 21, 1385-1400.

N. A. CAMPBELL (1982), Robust Procedures in Multivariate Analysis : Robust Canonical Variate Analysis. *Applied Statistics*, 31, 1-8.

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

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

M. L. EATON & D. TYLER (1994), The asymptotic Distribution of Singular values with Applications to Canonical Correlations and Correspondance Analysis. *Journal of Multivariate Analysis*, 50, 238-264.

F. R. HAMPEL, E. M. RONCHETTI, P. J. ROUSSEEUW & W. A. STAHEL (1986), *Robust Statistics : The Approach Based in Influence Functions*. New York : John Wiley and Sons.

H. HOTELLING (1936), Relations between two sets of variates. *Biometrika*, 28, 321-377.

P. L. HSU (1941), On the Limiting Distribution of the Canonical Correlations. *Biometrika*, 32, 38-45.

R. A. JOHNSON & D. W. WICHERN (1992) *Applied Multivariate Statistical Analysis*. Prentice Hall International Editions.

G. KARNEL (1991), Robust Canonical Correlation and Correspondance Analysis. *The Frontiers of Statistical Scientific and Industrial Applications*, (Volume II of the proceedings of ICOSCO-I, The First International Conference on Statistical Computing, 335-354).

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

H. P. LOPUHAÄ & P. J. ROUSSEEUW (1991), Breakdown Points of Affine Equivariant Estimators of Multivariate Location and Covariance Matrices. *The Annals of Statistics*, 19, 229-248.

R. A. MARONNA (1976), Robust M-estimators of Multivariate Location and Scatter. *The Annals of Statistics*, 4, 51-67.

N. J. PULLMAN (1976) *Matrix theory and its applications : selected topics*. Marcel Dekker, New York.

A. C. RENCHER (1998) *Multivariate Statistical Inference and Applications*, New York : Wiley.

M. ROMANAZZI (1992) Influence in Canonical Correlation Analysis. *Psychometrika*, 57, 237-259.

P. J. ROUSSEEUW (1985), Multivariate Estimation with Hight Breakdown Point. *Mathematical Statistics and Applications, Vol. B*, pp. 283-297, Dordrecht, the Netherlands : W. Grossmann, G. Pflug, I. Vincze, and W. Wertz.

P. J. ROUSSEEUW & K. VAN DRIESSEN (1999), A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41, 212-23.

P. J. ROUSSEEUW & B. C. VAN ZOMEREN (1990), Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association,* 85, 663-639.

R. SIBSON (1979), Studies in the Robustness of Multidimensional Scaling : Perturbational Analysis of Classical Scaling. *Journal of the Royal Statistical Society B*, 41, 217-229.

M. TENENHAUS (1998), *La Régression PLS. Théorie et Pratique*, Paris : Technip.