Identification d'observations aberrantes pour des distributions multivariées unimodales asymétriques et/ou à queues lourdes

Catherine Vermandele [1] & Vincenzo Verardi [2]

[1] *Université libre de Bruxelles, LMTD - CP 139, avenue F.D. Roosevelt 50, 1050 Bruxelles, Belgique,* vermande@ulb.ac.be

[2] *Université de Namur, Rempart de la Vierge 8, 5000 Namur, Belgique,* vverardi@unamur.be

Résumé. L'identification de valeurs extrêmes (aberrantes) s'avère particulièrement délicate en analyse multivariée lorsque la distribution sous-jacente est asymétrique et/ou à queues lourdes. Nous présenterons dans cette communication une méthode d'identification extrêmement simple, bien adaptée à ce type de distribution et qui n'exige qu'une faible complexité calculatoire. Cette méthode se fonde essentiellement sur la détermination d'une mesure spécifique du caractère extrême (aberrant) de chacune des observations multivariées étudiées, puis sur l'ajustement « robuste » (peu sensible à la présence de valeurs extrêmes), par une distribution dite de Tukey g-et-h, de la distribution des valeurs obtenues en appliquant une transformation monotone croissante fort simple à ces mesures du caractère extrême.

Mots-clés. Identification de valeurs extrêmes (aberrantes), distribution multivariée asymétrique, distribution multivariée à queues lourdes, distribution de Tukey g-et-h.

Abstract. In multivariate analysis, it is very difficult to identify outliers in case of skewed and/or heavy-tailed distributions. In this communication, we propose a very simple outlier identification tool that performs well with these types of distributions and that keeps the computational complexity low. The first step of the proposed method consists in allocating a specific outlyingness measure to each multivariate observation. The second step is to adjust in a "robust" way, using a Tukey g-and-h distribution, the distribution of the values obtained by applying a quite simple montone transformation on these outlyingness measures.

Keywords. Outlier identification, skewed multivariate distribution, heavy-tailed multivariate distribution, Tukey g-and-h distribution.

1 Introduction

L'identification de valeurs extrêmes (aberrantes) s'avère particulièrement délicate en analyse multivariée lorsque la distribution sous-jacente est asymétrique et/ou à queues

lourdes. Nous présenterons dans cette communication une méthode d'identification extrêmement simple, bien adaptée à ce type de distribution et qui n'exige qu'une faible complexité calculatoire.

La première étape de cette méthode consiste à affecter à chacune des observations multivariées de la série étudiée une mesure spécifique de son caractère extrême (aberrant) ; cette mesure est définie en adaptant quelque peu la mesure initialement proposée par Stahel (1981) et Donoho (1982), et « ajustée » en 2008 par Hubert et Van der Veeken. La seconde étape de la méthode — l'identification des observations extrêmes — se fonde sur l'ajustement « robuste » (c'est-à-dire peu sensible à la présence de valeurs extrêmes), par une distribution dite de Tukey g-et-h, de la distribution des valeurs obtenues en appliquant une transformation monotone croissante fort simple à ces mesures du caractère extrême.

2 La méthode « par projection » d'identification d'observations multivariées extrêmes (aberrantes)

Considérons un échantillon p-dimensionnel $\mathcal{X}^{(n)} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, avec $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^t$. Dans ce contexte multivarié, un point $\mathbf{x}_i$ est extrême (aberrant) s'il se situe loin de l'ensemble des observations dans au moins une direction de $\mathbb{R}^p$. Selon cette idée, Stahel (1981) et Donoho (1982) ont proposé de quantifier le caractère extrême d'une observation à partir de sa projection sur la direction de l'espace le long de laquelle cette observation est la plus extérieure. Plus précisément, étant donné une direction $\mathbf{a} \in \mathbb{R}^p$ telle que $\|\mathbf{a}\| = 1$, désignons par $\mathcal{X}_{\mathbf{a}}^{(n)} = \{\mathbf{x}_1^t\mathbf{a}, \ldots, \mathbf{x}_n^t\mathbf{a}\}$ la projection de l'ensemble $\mathcal{X}^{(n)}$ sur $\mathbf{a}$. Soient $\widehat{\mu}$ et $\widehat{\sigma}$, des statistiques univariées robustes de position et de dispersion (telles que la médiane et le MAD [1]), respectivement. Le caractère extrême par rapport à $\mathcal{X}^{(n)}$ de l'observation $\mathbf{x}_i$ le long de la direction $\mathbf{a}$ se mesure comme suit :

$$\mathrm{SDO}_{\mathbf{a}}(\mathbf{x}_i; \mathcal{X}^{(n)}) = \frac{|\mathbf{x}_i^t\mathbf{a} - \widehat{\mu}(\mathcal{X}_{\mathbf{a}}^{(n)})|}{\widehat{\sigma}(\mathcal{X}_{\mathbf{a}}^{(n)})}. \tag{1}$$

La *mesure (globale) de Stahel-Donoho du caractère extrême — Stahel-Donoho Outlyingness* — de $\mathbf{x}_i$ par rapport à $\mathcal{X}^{(n)}$ est alors donnée par

$$\mathrm{SDO}_i = \sup_{\mathbf{a} \in \mathcal{S}_p} \mathrm{SDO}_{\mathbf{a}}(\mathbf{x}_i; \mathcal{X}^{(n)}),$$

où $\mathcal{S}_p = \{\mathbf{a} \in \mathbb{R}^p : \|\mathbf{a}\| = 1\}$. Bien sûr, il nous faut nous contenter, en pratique, d'une valeur approchée de SDO_i en nous restreignant à considérer un ensemble fini $\widehat{\mathcal{S}}_p$ de directions sélectionnées aléatoirement [2].

1. MAD : Median Absolute Deviation
2. On considère le plus souvent $m = 250p$ directions différentes (cf. Maronna et Yohai, 1995).

Si les données $\mathbf{x}_i$ $(i = 1, \ldots, n)$ sont distribuées selon une loi normale p-variée, les mesures SDO_i $(i = 1, \ldots, n)$ sont distribuées asymptotiquement selon une loi χ^2_p (Maronna et Yohai, 1995). On peut alors identifier une observation $\mathbf{x}_i$ comme étant extrême si la mesure SDO_i qui lui correspond excède le quantile d'ordre $(1-\alpha)$ de la loi χ^2_p, où α est un niveau de probabilité faible préalablement fixé.

Cette procédure d'identification d'observations multivariées extrêmes souffre de deux problèmes majeurs : (i) si les données $\mathbf{x}_i$ $(i = 1, \ldots, n)$ ne sont pas gaussiennes, la distribution des mesures SDO_i est généralement inconnue (mais, typiquement, présente une asymétrie droite) et la règle d'identification des observations extrêmes fondée sur le quantile $\chi^2_{p;1-\alpha}$ risque d'être invalide ; (ii) la mesure du caractère extrême définie en (1) ne tient pas compte du caractère asymétrique que peut présenter la distribution de l'ensemble $\mathcal{X}^{(n)}_{\mathbf{a}}$ des projections des observations $\mathbf{x}_i$ sur $\mathbf{a}$, et n'est donc bien adaptée qu'à des données $\mathbf{x}_i$ de distribution symétrique elliptique. C'est pour pallier ces deux problèmes que Hubert et Van der Veeken (2008) ont proposé la méthode *ajustée* décrite ci-dessous.

3 La méthode *ajustée* de Hubert et Van der Veeken

La mesure du caractère extrême d'une observation proposée par Hubert et Van der Veeken fait intervenir les extrémités des moustaches de la boîte à moustaches *ajustée* introduite par Hubert et Vandervieren (2008). Pour la série statistique univariée $\{y_1, \ldots, y_n\}$, ces extrémités correspondent aux bornes de l'intervalle

$$\begin{array}{ll} \left[Q_{0.25} - 1.5e^{-4\mathrm{MC}}\mathrm{IQR}; Q_{0.75} + 1.5e^{3\mathrm{MC}}\mathrm{IQR}\right] & \text{if MC} \geq 0 \\ \left[Q_{0.25} - 1.5e^{-3\mathrm{MC}}\mathrm{IQR}; Q_{0.75} + 1.5e^{4\mathrm{MC}}\mathrm{IQR}\right] & \text{if MC} < 0, \end{array}$$

où $Q_{0.25}$ et $Q_{0.75}$ sont les 1$^{\mathrm{er}}$ et 3$^{\mathrm{e}}$ quartiles de la série $\{y_1, \ldots, y_n\}$, $\mathrm{IQR} = Q_{0.75} - Q_{0.25}$ est l'écart interquartile et MC est le *medcouple* — une mesure robuste de l'asymétrie, comprise entre -1 et $+1$ — de la série. Ces expressions des extrémités *ajustées* de la boîte à moustaches ont été déterminées en simulant une large variété de distributions asymétriques et en recherchant l'intervalle en dehors duquel ne se trouvent que 0.7% des observations en l'absence de contamination par des valeurs aberrantes.

Hubert et Van der Veeken (2008) définissent la mesure *ajustée* du caractère extrême de l'observation $\mathbf{x}_i$ le long de la direction $\mathbf{a}$ comme suit :

$$\mathrm{AO}_{\mathbf{a}}(\mathbf{x}_i; \mathcal{X}^{(n)}) = \begin{cases} \frac{\mathbf{x}_i^t\mathbf{a} - Q_{0.5}(\mathcal{X}^{(n)}_{\mathbf{a}})}{u_2(\mathcal{X}^{(n)}_{\mathbf{a}}) - Q_{0.5}(\mathcal{X}^{(n)}_{\mathbf{a}})} & \text{if } \mathbf{x}_i^t\mathbf{a} \geq Q_{0.5}(\mathcal{X}^{(n)}_{\mathbf{a}}) \\ \frac{Q_{0.5}(\mathcal{X}^{(n)}_{\mathbf{a}}) - \mathbf{x}_i^t\mathbf{a}}{Q_{0.5}(\mathcal{X}^{(n)}_{\mathbf{a}}) - u_1(\mathcal{X}^{(n)}_{\mathbf{a}})} & \text{if } \mathbf{x}_i^t\mathbf{a} < Q_{0.5}(\mathcal{X}^{(n)}_{\mathbf{a}}) \end{cases},$$

où $Q_{0.5}(\mathcal{X}^{(n)}_{\mathbf{a}})$ est la médiane de l'ensemble $\mathcal{X}^{(n)}_{\mathbf{a}}$ des projections, et $u_1(\mathcal{X}^{(n)}_{\mathbf{a}})$ et $u_2(\mathcal{X}^{(n)}_{\mathbf{a}})$ sont les extrémités de la moustache gauche et de la moustache droite, respectivement, de la boîte à moustaches ajustée associée à $\mathcal{X}^{(n)}_{\mathbf{a}}$. Cette définition prend en compte le fait

que la distribution des observations projetées $\mathbf{x}_i^t\mathbf{a}$ peut être asymétrique, ce qui induit alors une différence d'échelle entre la partie de la distribution à gauche de la médiane et la partie à droite de cette dernière.

La *mesure (globale) ajustée du caractère extrême — Adjusted Outlyingness —* de $\mathbf{x}_i$ par rapport à $\mathcal{X}^{(n)}$ est tout simplement donnée par

$$\mathrm{AO}_i = \sup_{\mathbf{a}\in\widehat{\mathcal{S}}_p} \mathrm{AO}_{\mathbf{a}}(\mathbf{x}_i; \mathcal{X}^{(n)}).$$

Hubert et Van der veeken (2008) suggèrent alors de déclarer comme extrême une observation multivariée $\mathbf{x}_i$ lorsque la mesure AO_i qui lui correspond est supérieure à l'extrémité de la moustache droite de la boîte à moustaches *ajustée* de Hubert et Vandervieren associée à la série $\{\mathrm{AO}_1, \ldots, \mathrm{AO}_n\}$.

L'approche de Hubert et Van der Veeken a l'avantage de ne plus devoir présupposer une distribution symétrique particulière pour les observations $\mathbf{x}_i$; seule l'unimodalité est *a priori* requise. Toutefois, elle souffre également de défauts non négligeables : (i) elle occasionne une complexité calculatoire substantielle (de l'ordre de $O(np\log n)$) liée à la nécessité de calculer le *medcouple* de $\mathcal{X}_{\mathbf{a}}^{(n)}$ pour chacune des $m = 250p$ directions $\mathbf{a} \in \mathbb{R}^p$ considérées ; (ii) le poucentage théorique d'observations identifiées comme extrêmes, en l'absence de contamination par des observations aberrantes, est fixé une fois pour toute à 0.7% ; (iii) ce pourcentage théorique risque de ne pas être respecté si la distribution des mesures AO_i $(i = 1, \ldots, n)$ s'avère sévèrement asymétrique et/ou possède une queue droite fort lourde.

La nouvelle méthode proposée apporte une solution à chacun de ces trois problèmes.

4 La distribution de Tukey g-et-h

A la fin des années 70', Tukey (1977) a introduit une nouvelle famille de distributions — appelées les distributions de Tukey g-et-h — définies à partir de transformations élémentaires de la distribution normale centrée réduite. Considérons, pour g et $h \in \mathbb{R}$, la fonction strictement monotone $\tau_{g,h}(\cdot)$ définie sur $\mathbb{R}$ comme suit : pour $g \neq 0$,

$$\tau_{g,h}(z) = \frac{1}{g}[\exp(gz) - 1]\exp(hz^2/2)$$

et, pour $g = 0$,

$$\tau_{0,h}(z) = \lim_{g\to 0} \tau_{g,h}(z) = z\exp(hz^2/2).$$

Soit Z une variable aléatoire de loi $N(0,1)$: pour $A \in \mathbb{R}$ et $B \in \mathbb{R}_0^+$, la variable aléatoire Y, définie par la transformation $Y = A + B\tau_{g,h}(Z)$, suit une distribution de Tukey g-et-h, de paramètre de position A et de paramètre d'échelle B ($Y \sim T_{g,h}(A,B)$). Le paramètre g

contrôle le sens et le degré de l'asymétrie [3], tandis que h contrôle l'épaisseur (ou élongation) de la distribution de Y. Les distributions $T_{g,h}(A,B)$ permettent de bien ajsuter une très large variété de distributions standards.

Différentes procédures pour l'estimation des paramètres de la distribution $T_{g,h}(A,B)$ ont été proposées dans la littérature. Nous proposons ici de faire appel à des estimateurs très simples de ces paramètres, définis exclusivement à partir des quantiles d'ordre 0.10, 0.25, 0.5, 0.75 et 0.90, et jouissant ainsi d'une bonne robustesse (point de rupture [4] égal à 10%).

5 La nouvelle méthode proposée

5.1 La mesure *asymétrique* du caractère extrême d'une observation

Comme Hubert et Van der Veeken (2008), nous proposons de modifier la mesure du caractère extrême d'une observation $\mathrm{SDO}_{\mathbf{a}}(\mathbf{x}_i;\mathcal{X}^{(n)})$ de Stahel-Donoho en prenant en compte l'asymétrie éventuelle de la distribution des points projetés $\mathbf{x}_i^t\mathbf{a}$. Nous définissons la mesure *asymétrique* du caractère extrême de l'observation $\mathbf{x}_i$ le long de la direction $\mathbf{a}$ comme suit :

$$\mathrm{ASO}_{\mathbf{a}}(\mathbf{x}_i;\mathcal{X}^{(n)}) = \begin{cases} \frac{\mathbf{x}_i^t\mathbf{a} - Q_{0.5}(\mathcal{X}_{\mathbf{a}}^{(n)})}{2c\left[Q_{0.75}(\mathcal{X}_{\mathbf{a}}^{(n)}) - Q_{0.5}(\mathcal{X}_{\mathbf{a}}^{(n)})\right]} & \text{if } \mathbf{x}_i^t\mathbf{a} \geq Q_{0.5}(\mathcal{X}_{\mathbf{a}}^{(n)}) \\ \frac{Q_{0.5}(\mathcal{X}_{\mathbf{a}}^{(n)}) - \mathbf{x}_i^t\mathbf{a}}{2c\left[Q_{0.5}(\mathcal{X}_{\mathbf{a}}^{(n)}) - Q_{0.25}(\mathcal{X}_{\mathbf{a}}^{(n)})\right]} & \text{if } \mathbf{x}_i^t\mathbf{a} < Q_{0.5}(\mathcal{X}_{\mathbf{a}}^{(n)}) \end{cases},$$

où $Q_{0.5}(\mathcal{X}_{\mathbf{a}}^{(n)})$, $Q_{0.25}(\mathcal{X}_{\mathbf{a}}^{(n)})$ et $Q_{0.75}(\mathcal{X}_{\mathbf{a}}^{(n)})$ sont, respectivement, la médiane, le 1$^{\text{er}}$ quartile et le 3$^{\text{e}}$ quartile de la série $\mathcal{X}_{\mathbf{a}}^{(n)}$ des projections, et $c = 1/(z_{0.75} - z_{0.25}) = 0.7413$ est une constante assurant, dans le cas gaussien, la convergence de l'estimateur d'échelle cIQR vers le paramètre d'échelle σ (l'écart-type). La *mesure (globale) asymétrique du caractère extrême — Asymmetrical Outlyingness —* de $\mathbf{x}_i$ par rapport à $\mathcal{X}^{(n)}$ est alors donnée par

$$\mathrm{ASO}_i = \sup_{\mathbf{a}\in\widehat{\mathcal{S}}_p} \mathrm{ASO}_{\mathbf{a}}(\mathbf{x}_i;\mathcal{X}^{(n)}).$$

5.2 La règle d'identification des observations extrêmes

La procédure proposée pour l'identification des observations extrêmes se décompose en quatre étapes :

3. $g = 0$ correspond à une distribution symétrique ; $g > 0$ (resp. $g < 0$) donne une distribution avec une asymétrie droite (resp. gauche).

4. Intuitivement, le point de rupture d'un estimateur est la proportion maximale d'observations aberrantes (arbitrairement grandes) auxquelles il peut faire face sans « se rompre » en prenant une valeur arbitraire. Plus le point de rupture d'un estimateur est élevé, plus ce dernier est robuste.

1. On standardise les mesures ASO_i afin d'obtenir de nouvelles valeurs appartenant à l'intervalle $(0,1)$: pour $i = 1, \ldots, n$, on calcule
$$\widetilde{\mathrm{ASO}}_i = \frac{\mathrm{ASO}_i}{\min_{1\leq j\leq n}(\mathrm{ASO}_j) + \max_{1\leq j\leq n}(\mathrm{ASO}_j)}.$$
2. On détermine, pour $i = 1, \ldots, n$, les valeurs $w_i = \Phi^{-1}\left(\widetilde{\mathrm{ASO}}_i\right)$, où $\Phi(\cdot)$ est la fonction de répartition de la loi $N(0,1)$.
3. On ajuste la distribution des valeurs w_i $(i = 1, \ldots, n)$ par une distribution de Tukey $T_{\widehat{g},\widehat{h}}(\widehat{A}, \widehat{B})$.
4. On détermine le quantile $\xi_{1-\alpha}$ d'ordre $1-\alpha$ $(\alpha \in (0, 0.5))$ de la distribution de Tukey $T_{\widehat{g},\widehat{h}}(\widehat{A}, \widehat{B})$ spécifiée à l'étape précédente, où α correspond au taux désiré de détection de valeurs extrêmes en l'absence de contamination des données par des observations aberrantes. Soit $\mathcal{I} = \{i = 1, \ldots, n | w_i > \xi_{1-\alpha}\}$; les observations $\mathbf{x}_i$ telles que $i \in \mathcal{I}$ sont alors identifiées comme étant des observations extrêmes dans l'ensemble de données $\mathcal{X}^{(n)}$.

Au cours de la communication, nous mettrons en avant la logique qui sous-tend les différentes étapes de la procédure décrite ci-dessus. Nous montrerons ensuite, à l'aide de quelques résultats empiriques et de simulations, que cette nouvelle procédure, malgré sa grande simplicité et son très faible coût en temps de calcul, présente de bonnes performances pour une très large variété—en termes de degré d'asymétrie et de lourdeur de queues—de distributions multivariées unimodales des observations $\mathbf{x}_i$.

Bibliographie

[1] Donoho, D. (1982), Breakdown properties of multivariate location estimators, Technical report, Harvard University, Boston, Qualifying paper.

[2] Hubert, M. and Van der Veeken, S. (2008), Outlier detection for skewed data, *J. Chemometrics*, 22(3-4), 235–246.

[3] Hubert, M. and Vandervieren, E. (2008), An adjusted boxplot for skewed distributions, *Comput. Stat. Data Anal.*, 52(12), 5186–5201.

[4] Maronna, R. and Yohai, V. (1995), The behavior of the Stahel-Donoho robust multivariate estimator, *Journal of the American Statistical Association*, 90(429), 330–341.

[5] Stahel, W. (1981), *Robuste Schätzungen : Infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen*, PhD thesis, ETH Zürich.

[6] Tukey, J. (1977), Modern techniques in data analysis. In *Proceedings of the NSF-Sponsored Regional Research Conference*.

[7] Verardi, V. and Vermandele, C. (2016), Outlier identification for skewed and/or heavy-tailed unimodal mutlivariate distributions, *Journal de la Société Française de Statistique*, 157(2), 90–114.