Robustesse de la profondeur scatter

Gaëtan Louvet [1] & Germain Van Bever [2]

[1] *Université libre de Bruxelles, Belgique, gaetan.louvet@ulb.be*
[2] *Université libre de Bruxelles, Belgique, germain.van.bever@ulb.be*

Résumé. La profondeur statistique fournit des outils non paramétriques robustes pour analyser les distributions. En effet, les fonctions de profondeur mesurent l'adéquation entre les paramètres d'une distribution et les mesures de probabilité sous-jacentes. Dans le cas *position*, un exemple de telle mesure est la profondeur de demi-espace de Tukey, dont les propriétés de robustesse ont déjà été largement étudiées. Récemment, des notions de profondeur pour les paramètres de *dispersion* ont été définies et étudiées. Les propriétés de robustesse de ces fonctions de profondeur scatter restent toutefois largement inconnues. Dans cet exposé, nous présentons différents résultats concernant la robustesse de la profondeur scatter ainsi que de la médiane associée. Nous en dérivons la fonction d'influence et le point de rupture, et déduisons la distribution asymptotique des versions empiriques de celles-ci.

Mots-clés. Profondeur scatter, robustesse, fonction d'influence, point de rupture, distribution asymptotique.

Abstract. Statistical depth provides robust nonparametric tools to analyze distributions. Depth functions indeed measure the adequacy of distributional parameters to underlying probability measures. In the *location* case, the celebrated halfspace depth has been widely studied, and its robustness properties are amply discussed. Recently, depth notions for *scatter* parameters have been defined and studied. The robustness properties of this latter depth function remain, however, largely unknown. In this talk, we present several results regarding the scatter depth function and its associated scatter median, including the influence function, the breakdown point and the asymptotic distribution.

Keywords. Scatter depth, robustness, influence function, breakdown point, asymptotic distribution.

1 Profondeur scatter

Génériquement, une fonction de profondeur $D(\cdot, \mathrm{P}) : \theta \mapsto \mathrm{D}(\theta, \mathrm{P})$ associe à tout paramètre θ une mesure de l'adéquation de celui-ci pour la distribution de probabilité P sous-jacente. Cet outil nonparamétrique permet typiquement de construire des méthodologies robustes pour l'estimation ou le test dans des distributions paramétriques.

Dans le cas *position*, la profondeur de demi-espace de Tukey (1975) est définie comme

$$HD_{\mathrm{loc}}(\theta, \mathrm{P}) = \inf_{||u||=1} \mathrm{P}\left[u'(X - \theta) \geq 0\right].$$

Ses propriétés de robustesse ont déjà été largement étudiées: fonction d'influence (Romanazzi (2001)), point de rupture (Liu *et al.* (2017)) et comportement asymptotique (Massé (2002)).

Récemment, des notions de profondeur pour les paramètres de *dispersion* ont été définies et étudiées. Les propriétés de robustesse de ces fonctions de profondeur *scatter* restent toutefois largement inconnues. Dans ce document, nous examinons principalement la robustesse de la profondeur du demi-espace de scatter introduite par Chen *et al.* (2018), dont l'expression est donnée par

$$HD_{\text{sc}}(\Sigma, \text{P}) = \inf_{||u||=1} \min\left(\text{P}\left[|u'(X - T_{\text{P}})| \leq \sqrt{u'\Sigma u}\right], \text{P}\left[|u'(X - T_{\text{P}})| \geq \sqrt{u'\Sigma u}\right]\right),$$

et où T_{P} est un estimateur de position. Nous considérons également la médiane de scatter associée, définie, pour $\mathcal{M}^p$ l'ensemble des matrices symétriques définies positives, comme

$$S_{HD}(\text{P}) = \text{argmax}_{\Sigma \in \mathcal{M}^p} HD_{\text{sc}}(\Sigma, \text{P}).$$

2 Fonction d'influence

La fonction d'influence mesure la sensibilité d'une fonctionnelle S à l'ajout d'une faible contamination sur la distribution P. Elle se définit formellement comme

$$\text{IF}(z; S(\text{P})) = \lim_{\varepsilon \to 0^+} \frac{S(\text{P}_{\varepsilon,z}) - S(\text{P})}{\varepsilon} = \frac{\partial}{\partial \varepsilon} S(\text{P}_{\varepsilon,z})\Big|_{\varepsilon=0^+},$$

où $\text{P}_{\varepsilon,z} = (1-\varepsilon)\text{P} + \varepsilon\Delta(z)$ correspond à la distribution contaminée et $\Delta(z)$ est la distribution de Dirac en $z \in \mathbb{R}^p$. Les résultats obtenus concernant la fonction d'influence de la profondeur scatter se distinguent selon l'estimateur de position utilisé et/ou les hypothèses de continuité sur la distribution sous-jacente. Nous décrivons ceux-ci de manière informelle ci-dessous.

- Dans le cas où le paramètre de position est supposé connu, la fonction d'influence de la profondeur scatter est bornée et s'écrit comme

  $$\text{IF}(z; HD_{\text{sc}}(\Sigma, \text{P})) = I[z \in A] - HD_{\text{sc}}(\Sigma, \text{P}),$$

  pour $A = A(\text{P})$ un ensemble dépendant de la distribution P.

- Pour des distributions absolument continues et un estimateur de position T_{P} quelconque, nous trouvons, sous certaines hypothèses sur P et T_{P}, pour une certaine fonction g, bornée si $\text{IF}(.; T_P)$ est bornée, que

  $$-HD_{\text{sc}}(\Sigma, \text{P}) + g(\text{IF}(z; T_{\text{P}})) \leq \text{IF}(z; HD_{\text{sc}}(\Sigma, \text{P})) \leq 1 - HD_{\text{sc}}(\Sigma, \text{P}) + g(\text{IF}(z; T_{\text{P}})).$$

- Dans le cas de distributions discrètes, il est impossible de dégager une expression générale explicite pour la fonction d'influence. Dans le cas où nous choisissons la médiane de demi-espace et que cette dernière est unique, nous retrouvons

  $$\text{IF}(z; HD_{\text{sc}}(\Sigma, \text{P})) = I[z \in A] - HD_{\text{sc}}(\Sigma, \text{P}).$$

L'obtention de la fonction d'influence de la médiane scatter en toute généralité est également impossible. Cependant, dans le cas de distributions elliptiques de densité radiale f et de matrice de dispersion Σ_*, nous trouvons, pour certaines valeurs de z,

$$\mathrm{IF}(z; S_{HD}(\mathrm{P})) = -\Sigma_*/(2f(1)).$$

Pour des distributions plus générales, il est également possible d'obtenir des bornes sur la fonction d'influence.

3 Point de rupture

Le point de rupture d'une fonctionnelle est informellement défini comme le nombre d'observations à modifier avant que celle-ci n'atteigne sa ou ses bornes. Dans le cas de la profondeur, nous considérons donc, pour une jeu de données X, les points de rupture *par substitution*

$$\mathrm{BP}(HD_{\mathrm{sc}}(\Sigma, X)) = \min_{1\leq m\leq n}\{\frac{m}{n} : \sup_{X^m} HD_{\mathrm{sc}}(\Sigma; X^m) = 0\},$$

et

$$\mathrm{BP}(S_{HD}(X)) = \min(\mathrm{BP}^0(S_{HD}(X)), \mathrm{BP}^\infty(S_{HD}(X))),$$

où

$$\mathrm{BP}^0(S_{HD}(X)) = \min_{1\leq m\leq n}\{\frac{m}{n} : \inf_{X^m} \lambda_p(S_{HD}(X^m)) = 0\},$$

et

$$\mathrm{BP}^\infty(S_{HD}(X)) = \min_{1\leq m\leq n}\{\frac{m}{n} : \sup_{X^m} \lambda_1(S_{HD}(X^m)) = \infty\}$$

et où X^m représente X où m données ont été arbitrairement modifiées et où $\lambda_1(S)$ et $\lambda_p(S)$ désignent, respectivement, la plus grande et la plus petite valeur propre de la matrice S. Les résultats sont les suivants:

- Lorsque l'estimateur de position T_P est connu et fixé, la définition de profondeur implique que $\mathrm{BP}(HD_{\mathrm{sc}}(\Sigma, X)) = HD_{\mathrm{sc}}(\Sigma, X)$. Le cas position non fixé dépendra de la fonctionnelle de position utilisée.

- En notant $M = n \max_\Sigma HD_{\mathrm{sc}}(\Sigma, X)$, nous trouvons comme borne inférieure

  $$\mathrm{BP}^0(S_{HD}, X) \geq (\frac{\lceil M/2\rceil - p}{n})^+ \qquad \text{et} \qquad \mathrm{BP}^\infty(S_{HD}, X) \geq \frac{\lceil M/2\rceil}{n}.$$

  Asymptotiquement, le point de rupture atteint donc, dans le pire des cas, une valeur de $\max_\Sigma HD(\Sigma, \mathrm{P})/2$ ($= 0.25$ pour des distributions elliptiques par exemple).

4 Distribution asymptotique

Enfin, nous dérivons la distribution asymptotique de la médiane scatter pour des distributions continues P. Notons Σ_0 la médiane scatter sous P. Soit G un vecteur aléatoire de même distribution que la distribution asymptotique de $\sqrt{n}(T_n - T_0)$. Soit $x : S^{p-1} \to \mathbb{R}$ un processus stochastique défini sur la sphère unité par, pour tout $u \in S^{p-1}$, $x(u) = K_1(u)u'G + \nu_F A_i(u, \Sigma_0, T_0)$, où A_i un sous espace et ν_F un pont brownien. Sous certaines hypothèses sur P et pour certains sous-ensemble $V^i(0)$ et $V^o(0)$ de S^{p-1}, nous trouvons

$$\sqrt{n}(\Sigma_n - \Sigma_0) \to \arg\max_{\Sigma} \min(\inf_{V^i(0)} K_2(u)u'\Sigma u + x(u), \inf_{V^o(0)} -(K_2(u)u'\Sigma u + x(u))),$$

à condition que le maximum de la fonction ci-dessus est unique.

Bibliographie

Tukey, J.W. (1975) Mathematics and the picturing of data. In: Proceedings of the International Congress of Mathematicians, pp. 523-531

Romanazzi, M. (2001) Influence function of halfspace depth. *J. Multivariate Anal.* 77(1), pp. 138-161

Massé, J.-C. (2002) Asymptotics for the Tukey Median. *J. Multivariate Anal.* 81(1), pp. 286-300

Liu, X., Zuo, Y. et Wang, Q. (2017) Finite sample breakdown point of Tukey's halfspace median. *Sci. China Math.* 60, pp. 861-874

Chen, M., Gao, C. et Ren, Z. (2018) Robust covariance and scatter matrix estimation under Huber's contamination model. *Ann. Statist.* 46(5), pp. 1932-1960