De meilleures régions de confiance pour les distributions à symétrie sphérique

Christian Robert et George Casella

BU-993-M October 1988

Résumé – Lorsqu'on recentre la région de confiance usuelle en un estimateur de James-Stein, on peut, sous certaines conditions, obtenir une probabilité d'occurence uniformément supérieure. Nous étendons ce résultat, déjà connu pour l'estimation de la moyenne d'un vecteur normal, à l'estimation du paramètre de position de lois à symétrie sphérique lorsque le paramètre d'échelle est inconnu.

Improved confidence sets for spherically symmetric distributions

Abstract – Domination of the least squares estimators by classes of shrinkage estimators has been intensively studied. When the risk is related to the coverage probability of the recentered confidence set, Hwang and Casella (1982) have established the domination of the lse by the positive-part James Stein estimators (1.1) for a range of values of the constant a and Hwang and Chen (1986) have generalized this result for spherically symmetric distributions. However, the case of unknown scale parameters has not been dealt with analytically. We give here a class of recentered set estimators which dominate the usual one for a class of ssd; the constant a belongs to an interval $[0,a_0]$ with upper bound solution of (2.3).

1. <u>Introduction</u> – Depuis les travaux initiaux de Stein (1956), l'estimation du paramètre de position des lois à symètrie sphérique a été étudiée par de nombreux auteurs (on pourra consulter Berger (1985) pour une revue). La plupart des travaux ne considérent que l'estimation de la moyenne d'une loi normale multidimensionnelle mais Cellier, Fourdrinier et Robert (1987) ont montré que, si l'on disposait d'observations en nombre strictement supérieur au nombre de paramètres estimés (situation qui inclut le cas de la variance connue à un facteur multiplicatif près), les estimateurs à rétrécisseur sont robustes; les conditions suffisantes de domination de l'estimateur des moindre carrés obtenues pour la loi normale peuvent être étendues aux lois à symétrie sphèrique et ne dépendent pas de la loi considérée.

Si l'on observe $z = (x',y')'$, résultant d'une loi à symétrie sphèrique de dimension $p + q$, de paramètre de position $(\theta',0')'$ et de matrice de dispersion $\sigma^2 I_{p+q}$, il est bien connu que les estimateurs de James-Stein tronqués

$$\delta_a^+(z) = \left(1 - a\frac{\|y\|^2}{\|x\|^2}\right)^+ x \, , \tag{1.1}$$

où $(t)^+ = \max(t,0)$, dominent uniformément l'estimateur des moindres carrés, $\delta_0(z) = x$, pour le risque quadratique usuel si $0 \leq a \leq 2(p\text{-}2)\frac{2(p\text{-}2)}{q+2}$. L'extension au cas d'une loi à symètrie sphérique découle de Cellier *et al.* Les estimateurs de James-Stein tronqués ne sont pas admissibles mais on ne connait pas d'estimateur les dominant uniformément (voir Bock (1988)).

Ces estimateurs présentent également un intérêt pour l'estimation par région de confiance. En effet, Hwang et Casella (1982) ont montré que la région de confiance usuelle,

$$C^0 = \{\theta; \|\theta\text{-}x\|^2 \leq c\} \tag{1.2}$$

est dominée par la région de confiance recentrée

$$C_a^0 = \left\{\theta; \|\theta - \left(1\text{-}a\frac{1}{\|x\|^2}\right)^+ x\|^2 \leq c\right\} . \tag{1.3}$$

lorsque $a \in [0,a_0]$ et x suit une loi normale, $\eta(\theta,I_p)$, au sens où ces deux régions ont même volume et

$$p_\theta(\theta \in C^0) \leq p_\theta(\theta \in C^0_a) \ ,$$

uniformément en θ. Ils ont ensuite élargi la borne a_0 (Hwang et Casella, 1984) et étendu leur résultat à une classe plus grande d'estimateurs (Casella et Hwang, 1987). La borne a_0 dépend du rayon c (i.e., du degré de confiance choisi) et décroit vers o quand c tend vers $+\infty$.

Hwang et Chen (1986) ont donné des bornes pour les lois à symétrie sphérique, les bornes dépendant explicitement de la distribution. Dans tous les cas, la matrice de dispersion est totalement connue et le seul résultat pour le cas de la variance connue à un facteur multiplicatif près est asymptotique (Chen et Hwang, 1988).

Le résultat que nous avons obtenu permet de proposer des régions de confiance améliorées pour des lois à symétrie sphérique dont le paramètre d'échelle est inconnu (l'extension à des lois à symétrie elliptique est immédiate). Malheureusement, la classe considérée ne comprend pas la loi normale, pour des raisons que nous exposerons dans la section 3. Plus fondamentalement, il apparaît que l'équivalent du résultat ponctuel, à savoir la robustesse des estimateurs à rétrécisseur, n'existe pas dans ce cadre, puisque les bornes sur a dépendent également de la distribution des observations.

2. Domination de la région de confiance usuelle – Dans le cas normal, il est bien connu que la région de confiance usuelle au niveau α $(0 \leq \alpha \leq 1)$ s'écrit

$$C^0 = \{\theta;\ ||\theta\text{-}x||^2 \leq k^2||y||^2\}\ , \tag{2.1}$$

où k^2 est le produit par $\frac{p}{q}$ du $(1\text{-}\alpha)$ - quantile d'une loi de Fisher à p et q degrés de liberté. Il découle en fait du Théorème 11 de Kelker (1970) que ce résultat est vrai pour toutes les lois à symétrie sphérique.

Considérant la région recentrée,

$$C_a = \{\theta;\ ||\theta\text{-}\delta_a^+(z)||^2 \leq k^2||y||^2\}\ , \tag{2.2}$$

nous donnons ci-dessous des conditions suffisantes (sur a) pour que C_a domine C^0 (au sens de la probabilité de ces ensembles). Notons f la fonction de $\mathbb{R}$ dans $\mathbb{R}$ telle que la densité de z s'écrive $f\Big((||x\text{-}\theta||^2 + ||y||^2)/\sigma^2\Big)$, où σ^2 est le paramètre d'echelle inconnu. Alors

Théorème 1 Si a_0 est solution de

$$\inf_w \left\{ w^2 \inf_{\alpha_0 \leq v \leq \alpha_1} \frac{f'(v+w^2)}{f(v+w^2)} \right\} = -\ \frac{p\text{-}2}{2k\sqrt{a}}\ \mathrm{Ln}\left(\frac{k + \sqrt{k^2+a}}{\sqrt{a}} \right), \tag{2.3}$$

où $\alpha_0 = (k - \sqrt{a})^2 w^2$ et $\alpha_1 = (k^2+a)w^2$,

pour tout à ϵ $[0,a_0]$, C_a domine C^0.

et, dans un cas plus particulier,

Théorème 2 Si

(a) $g = \mathrm{Ln}(f)$ est convexe,

(b) l'ensemble

$$\left\{ t \geq 1;\ \frac{p\text{-}2}{t} + as^2 g'\{s^2 a(t\text{-}t^{-1}) + [1 + (k\text{-}\sqrt{a})^2]s^2\}(1+t^2) \geq 0 \right\}$$

est un intervalle pour tout s,

(c) $g\{as^2(t^*-t^{*-1}) + [1+(k-\sqrt{a})^2]s^2\} - g\{[1+(k-\sqrt{a})^2]s^2\} \geq -(p-2)\mathrm{Ln}(t^*)$,

pour tout s, où $t^* = (k+\sqrt{k^2+a})/\sqrt{a}$,

C_a domine C^0 uniformément en θ.

Preuves Se déduisent de Hwang et Chen (1986) en travaillant conditionnellement à $||y||^2$.

3. Conséquences – Bien que ces résultats soient énoncés pour une loi à symètrie sphérique quelconque, ils n'admettent de solutions non triviales (i.e., $a \neq 0$) que pour certaines distributions. En particulier, ils ne sont d'aucune utilité dans le cadre de la loi normale [$f(t) \propto \exp(-t/2)$] ou double-exponentielle [$f(t) \propto \exp(-\sqrt{t})$]. L'exclusion de la loi normale découle naturellement de Hwang (1985, Theorem 3.10) car il établit qu'il n'existe pas de région de confiance dominant uniformément C^0 pour chaque degré de confiance dans le cas d'une loi normale, ce qui implique, pour notre problème, qu'une approche travaillant conditionnellement à $||y||^2$ est vouée à l'échec.

Au contraire, Hwang (1985) a montré qu'une solution est possible dans le cas d'une loi de Student multivariée

$$\left[f(t) \propto \left(1 + \frac{t}{N}\right)^{-(N+p+q)/2} \right]$$

et il apparaît en fait que les théorèmes 1 et 2 conduisent à des intervalles $[0,a_0]$ (de possibles constantes a) qui ne sont pas réduits à $\{0\}$. Etant donné que le théorème 2 conduit à des bornes plus grandes (voir Hwang et Chen, 1986), nous avons

Corollaire 1 Si q_0 est solution de

$$\frac{N+p+q}{2(p-2)} \mathrm{Ln}\left(1 + \sqrt{a}\,\frac{(k+\sqrt{k^2+a})^2 - a}{(k+\sqrt{k^2+a})[1+(k-\sqrt{a})^2]}\right) = \mathrm{Ln}\left(\frac{k+\sqrt{k^2+a}}{\sqrt{a}}\right), \tag{3.1}$$

C_a domine uniformément C^0 pour $a \in [0,a_0]$ dans le cas d'une loi de Student multivariée à N degrés de liberté.

Bien entendu, la solution de (3.1) tend vers 0 quand N tend vers $+\infty$, conformément au résultat pour la loi normale. La table ci-dessous présente les solutions de (3.1) pour un degré de confiance de 5% et N=5:

p \ q	5	10	20
5	0.58	0.21	0.10
10	2.03	0.68	0.30
20	5.42	1.73	0.73
30	8.98	2.81	1.17
60	19.84	6.05	2.42

4. Conclusion – Les bornes de la table précédente sont inférieures aux bornes obtenues pour l'estimation ponctuelle (voir Section 1) et par Hwang et Chen (1986). Elles indiquent que la méthode de conditionnement employée est relativement grossiére et que les améliorations futures devront utiliser des techniques permettant l'intégration par rapport à $||y||^2$, afin d'inclure également la loi normale. Cette exclusion de la loi normale est certes un inconvénient de la méthode précédente mais on peut noter que, même pour des valeurs modérées du nombre de degrés de liberté, la loi de Student est relativement proche de la loi normale et que, d'autre part, l'utilisation de la loi de Student offre une plus grande souplesse de modélisation (voir Zellner, 1976). Enfin, il semble certain, au vu de nombreuses simulations (voir Chen et Hwang, 1988), que le résultat peut être étendu aux lois normales, au prix d'un raffinement des méthodes de démonstration employées.

Références bibliographiques

[1] Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. Springer-Verlag.

[2] Bock, M.E. (1988). Shrinkage estimators: pseudo-Bayes rules for normal mean vectors. In *Statistical Decision Theory and Related Topics, IV, Vol.* 1, (S. Gupta, J. Berger, eds.), 281-297.

[3] Casella, G. and Hwang, J. (1987). Employing vague prior information in the construction of confidence sets. *J. Mult. Anal.* 21(1), 79-104.

[4] Cellier, D., Fourdrinier, D. et Robert, C. (1987). Estimateurs à rétrécisseur du paramètre de position d'une loi à symétrie sphérique. *Note aux C.R.A.S.* **304**, série 1, 439-442.

[5] Chen, J. and Hwang, J. (1988). Improved set estimators for the coefficients of a linear model when the error distribution is spherically symmetric with unknown variances. *Can. J. Statist.* 16(2) (à paraître).

[6] Hwang, J. (1985). Universal domination and stochastic domination: estimation simultaneously under a broad class of loss functions. *Ann. Stat.* 13(1), 295-314.

[7] Hwang, J. and Casella, G. (1982). Minimax confidence sets for the mean of a multivariate normal distribution. *Ann. Stat.* 10(3), 868-881.

[8] Hwang, J. and Casella, G. (1984). Improved set estimators for a multivariate normal mean. *Statist. Decision*, Supplement issue 1, 3-16.

[9] Hwang, J. and Chen, J. (1986). Improved confidence sets for the coefficients of a linear model with spherically symmetric errors. *Ann. Stat.* 14(2), 444-460.

[10] Kelker, D. (1970). Distribution theory of spherical distributions and a location-scale parameter generalization. *Sankhyā A* 32, 419-430.

[11] Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. 3rd Berkeley Symp. Math. Stat. Prob.* 1, 197-206.

[12] Zellner, A. (1976). Bayesian and non-Bayesian analysis of the regression model with multivariate Student-t error term. *JASA* 71, 400-405.

Department of Mathematics
12a White Hall
Cornell University
Ithaca, NY 14853
U.S.A.

Biometrics Unit
337 Warren Hall
Cornell University
Ithaca, NY 14853
U.S.A.