Estimation non-paramétrique des quantiles extrêmes conditionnels

Laurent Gardes, Stephane Girard, Alexandre Lekina

ESTIMATION NON-PARAMETRIQUE DES QUANTILES EXTREMES CONDITIONNELS

Laurent GARDES, Stéphane GIRARD & Alexandre LEKINA

Équipe Mistis, INRIA Rhône-Alpes & LJK
Inovallée, 655, av. de l'Europe, Montbonnot, 38334 Saint-Ismier cedex, France.
Laurent.Gardes@inria.fr, Stephane.Girard@inria.fr & Alexandre.Lekina@inria.fr

Résumé

Nous proposons dans le cas des distributions à queue lourde une méthode d'estimation des quantiles extrêmes en présence d'une covariable. La loi limite d'un tel estimateur est ensuite donnée en fonction de la vitesse de convergence de l'ordre du quantile vers un. Pour conclure, une illustration sur données simulées est présentée.

Abstract

We propose a method to estimate quantiles from heavy-tailed distributions when covariate information is available and in the case where the order of the quantile converges to one as the sample size increases. Asymptotic distribution of such an estimator is established in the case where the quantile is in the range of data or near and even beyond the sample. An illustration on simulated data is provided.

Mots-clés : Extrêmes, modèles semi et non paramétriques, quantiles conditionnels.

1 Contexte

Soit $Y \in \mathbb{R}$ une variable aléatoire associée à une covariable non-aléatoire $x \in E$, où E désigne un espace métrique (non nécessairement de dimension finie) muni d'une distance d. On note par $F(.,x)$ la fonction de répartition conditionnelle de Y sachant x et on suppose qu'elle admet un unique quantile conditionnel d'ordre $(1-\alpha)$ défini par,

$$F\left(q\left(\alpha,x\right),x\right)=1-\alpha,$$

pour tout $x \in E$ et $\alpha \in]0,1[$. On s'intéresse à l'estimation du réel $q(\alpha,x)$ lorsque la fonction de répartition conditionnelle $F(.,x)$ est dîte à queue lourde[1], i.e pour tout $\lambda > 0$,

$$\lim_{\alpha\to 0}\frac{q(\lambda\alpha,x)}{q(\alpha,x)}=\lambda^{-\gamma(x)}, \tag{1}$$

où $\gamma(.) > 0$ est une fonction inconnue de la covariable x appelée "indice des valeurs extrêmes conditionnel" (Gardes et Girard (2008)). On dit que le quantile conditionnel $q(.,x)$ est à variations régulières d'indice $-\gamma(x)$. On pourra se référer à Bingham,

[1]En statistique des valeurs extrêmes, on parle de domaine d'attraction de Fréchet.

Goldie et Teugels (1987) pour une explication détaillée de la théorie des fonctions à variations régulières.

Soit $\{(Y_i, x_i), i = 1, ..., n\}$ des observations indépendantes et de même loi que (Y, x), notre but est de construire en un point $t \in E$, un estimateur du quantile conditionnel $q(\alpha, t)$ lorsque α tend vers 0. Un exemple de problème pourrait être présenté comme suit.

Exemple *En fonction de la localisation géographique x, la hauteur d'eau d'un fleuve est modélisée par une variable aléatoire Y. On dispose de $Y_1, ..., Y_n$ hauteurs d'eau annuelles respectivement en n lieux $x_1, ..., x_n$ déterministes. Calculer pour une probabilité $p < 1/n$, une hauteur d'eau h extrême en un point t vérifiant $\mathbb{P}(Y > h|t) = p$.*

2 Définition des estimateurs

Nous utilisons une méthode de fenêtres mobiles pour construire notre estimateur. Pour cela, on introduit une boule centrée en t, de rayon $r > 0$, notée $B(t, r)$ et définie par

$$B(t, r) = \{x \in E, d(r, t) \leq r\}.$$

Etant donné $h_{n,t} = h_t$, une suite positive tendant vers 0 quand n tend vers l'infini, on se propose de ne sélectionner que les observations Y_i pour lesquelles les covariables x_i sont dans la boule $B(t, h_t)$. La proportion de tels points est ainsi donnée par

$$\varphi(h_t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{x_i \in B(t, h_t)\}$$

et joue un rôle central dans cette étude. De façon similaire à la notion de probabilité de petite boule utilisée en analyse fonctionnelle dans Ferraty et Vieu (2006), $\varphi(t)$ décrit comment cet ensemble de points se concentre dans un voisinage de t lorsque $h_t \to 0$. On note $\{Z_i(t), i = 1, ..., m_t\}$, les observations retenues par la procédure de sélection et on désigne par $Z_{1,m_t}(t) \leq ... \leq Z_{m_t,m_t}(t)$ les statistiques ordonnées correspondantes.

Dans cet article, nous nous focalisons sur l'estimation des quantiles "extrêmes" conditionnels d'ordre $1 - \alpha_{m_t}$. Ici on parle de quantile extrême si pour n tendant vers l'infini on a $\alpha_{m_t} \to 0$. En fonction de la vitesse de convergence de α_{m_t} vers 0, trois situations sont envisagées :

(S.1) α_{m_t} converge "lentement" vers 0, i.e $\alpha_{m_t} \to 0$ et $m_t \alpha_{m_t} \to \infty$. L'estimation du quantile extrême conditionnel requiert d'interpoler à l'intérieur de l'échantillon car $q(\alpha_{m_t}, t)$ est presque sûrement inférieur à l'observation maximale. On propose alors d'estimer $q(\alpha_{m_t}, t)$ par :

$$\hat{q}_1(\alpha_{m_t}, t) = Z_{m_t - \lfloor m_t \alpha_{m_t} \rfloor + 1, m_t}(t). \tag{2}$$

(S.2) α_{m_t} converge "rapidement" vers 0, i.e $\alpha_{m_t} \to 0$, $m_t\alpha_{m_t} \to c \in [1, \infty[$ et $\lfloor m_t\alpha_{m_t} \rfloor \to \lfloor c \rfloor$. Pour tout n assez grand, $\lfloor m_t\alpha_{m_t} \rfloor = c > 0$ et l'estimation du quantile extrême conditionnel repose sur les plus grandes observations situées au voisinage de la frontière de l'échantillon, mais toujours dans l'ensemble des données. Par conséquent, on peut réutiliser l'estimateur défini en (2).

(S.3) α_{m_t} converge "très rapidement" vers 0, i.e $\alpha_{m_t} \to 0$ et $m_t\alpha_{m_t} \to c \in [0, 1[$. Estimer le quantile extrême conditionnel nécessite d'extrapoler au-delà des observations puisque $q(\alpha_{m_t}, t)$ est supérieur à l'observation maximale avec probabilité $e^{-c} \geq e^{-1}$. Dans une telle situation, on propose d'adapter l'estimateur de Weissman (1978) au cas conditionnel. On estime alors $q(\alpha_{m_t}, t)$ par :

$$\hat{q}_2(\alpha_{m_t}, t) = \hat{q}_1(\beta_{m_t}, t) \, (\beta_{m_t}/\alpha_{m_t})^{\hat{\gamma}_n(t)} , \tag{3}$$

où β_{m_t} satisfait **(S.1)** et $\hat{\gamma}_n(t)$ est un estimateur de l'indice des valeurs extrêmes conditionnel. De tels estimateurs ont déjà été proposés par Beirlant et Goegebeur (2004) puis généralisés par Gardes et Girard (2008).

3 Lois asymptotiques

Il convient tout d'abord de donner quelques conditions et résultats auxiliaires utiles pour établir la loi asymptotique de nos estimateurs. Les démonstrations des résultats de ce paragraphe sont disponibles dans Gardes, Girard et Lekina (2008). Dans tout ce qui suit, on fixe t dans E et on suppose que :

(A) la fonction quantile conditionnel $\alpha \in]0, 1[\mapsto q(\alpha, t) \in]0, +\infty[$ est dérivable et la fonction définie par $\alpha \in]0, 1[\mapsto \Delta(\alpha, t) = \gamma(t) + \alpha \dfrac{\partial \log q}{\partial \alpha}(\alpha, t) \in]0, +\infty[$ est continue et telle que $\lim\limits_{\alpha \to 0} \Delta(\alpha, t) = 0$.

L'hypothèse **(A)** a pour but de contrôler le comportement de la fonction log-quantile quant à sa première variable. C'est une condition suffisante pour que la fonction de répartition conditionnelle $F(., x)$ soit à queue lourde (Bingham, Goldie et Teugels (1987), chap 1). En ce qui concerne sa seconde variable, la plus grande oscillation de la fonction log-quantile est définie pour tout $a \in]0, 1/2[$ par :

$$\omega_n(a) = \sup \left\{ \left| \log \frac{q(\alpha, x)}{q(\alpha, x')} \right| , \ \alpha \in]a, 1 - a[, \ (x, x') \in B(t, h_t)^2 \right\}.$$

Notre premier résultat est dédié à l'étude de la position du quantile extrême conditionnel dans l'ensemble des données.

Proposition 1 *Si pour tout $\delta > 0$, $(m_t\alpha_{m_t})^2 \omega_n(m_t^{-(1+\delta)}) \to 0$, alors*
- *sous* **(S.1)**, $\mathbb{P}(Z_{m_t, m_t} < q(\alpha_{m_t}, t)) \to 0$,
- *sous* **(S.2)** *ou* **(S.3)**, $\mathbb{P}(Z_{m_t, m_t} < q(\alpha_{m_t}, t)) \to e^{-c}$.

Les théorèmes suivants établissent la loi limite d'un estimateur de quantile extrême conditionnel construit à partir de notre procédure d'estimation.

Théorème 1 *Soit* (α_{m_t}) *une suite satisfaisant* **(S.1)**. *Si* $(m_t\alpha_{m_t})^2\omega_n(m_t^{-(1+\delta)}) \to 0$ *pour tout* $\delta > 0$, *alors*

$$(m_t\alpha_{m_t})^{1/2}\left(\frac{\hat{q}_1(\alpha_{m_t},t)}{q(\alpha_{m_t},t)}-1\right) \xrightarrow{d} \mathcal{N}\left(0,\gamma^2(t)\right).$$

Dans la situation **(S.1)**, la variance asymptotique étant inversement proportionnelle à α_{m_t}, l'estimation du quantile extrême conditionnel est d'autant plus stable qu'on s'éloigne de la frontière de l'échantillon.

Théorème 2 *Soit* (α_{m_t}) *une suite satisfaisant* **(S.2)**. *Si* $(m_t\alpha_{m_t})^2\omega_n(m_t^{-(1+\delta)}) \to 0$ *pour tout* $\delta > 0$, *alors*

$$\left(\frac{\hat{q}_1(\alpha_{m_t},t)}{q(\alpha_{m_t},t)}-1\right) \xrightarrow{d} \mathcal{E}\left(c,\gamma(t)\right)$$

où $\mathcal{E}\left(c,\gamma(t)\right)$ *est une loi non dégénérée.*

Dans la situation **(S.2)**, la loi asymptotique du quantile n'est pas gaussienne et son expression est assez compliquée. En outre, l'estimateur $\hat{q}_1(.,t)$ n'est pas consistant.

Théorème 3 *Soit* (β_{m_t}) *une suite satisfaisant* **(S.1)** *et soit* (α_{m_t}) *une suite telle que* $\alpha_{m_t} < \beta_{m_t}$. *On pose* $\zeta_{m_t} = (m_t\alpha_{m_t})^{1/2}\log\left(\beta_{m_t}/\alpha_{m_t}\right)$ *et pour tout* $b \in]0,1[$ *on definit* $\bar{\Delta}(b,t) = \sup_{\alpha\in]0,b[}|\Delta(\alpha,t)|$. *Si* $(m_t\alpha_{m_t})^2\omega_n(m_t^{-(1+\delta)}) \to 0$ *pour tout* $\delta > 0$ *et s'il existe une suite positive* $v_n(t)$ *et une loi* $\mathcal{D}$ *telle que*

$$v_n(t)\left(\hat{\gamma}_n(t)-\gamma(t)\right) \xrightarrow{d} \mathcal{D},$$

alors, deux situations se présentent :

(i) Soit la loi asymptotique découle de $\hat{q}_1(\beta_{m_t},t)$ *et sous la condition additionnelle*

$$\zeta_{m_t}\max\left\{v_n^{-1}(t),\bar{\Delta}(\beta_{m_t},t)\right\} \to 0,$$

nous avons

$$(m_t\alpha_{m_t})^{1/2}\left(\frac{\hat{q}_2(\alpha_{m_t},t)}{q(\alpha_{m_t},t)}-1\right) \xrightarrow{d} \mathcal{N}\left(0,\gamma^2(t)\right).$$

(ii) Sinon, elle provient de $\hat{\gamma}_n(t)$ *et sous la condition additionnelle*

$$v_n(t)\max\left\{\zeta_m^{-1}(t),\bar{\Delta}(\beta_{m_t},t)\right\} \to 0,$$

nous avons

$$\frac{v_n(t)}{\log\left(\beta_{m_t}/\alpha_{m_t}\right)}\left(\frac{\hat{q}_2(\alpha_{m_t},t)}{q(\alpha_{m_t},t)}-1\right) \xrightarrow{d} \mathcal{D}.$$

La loi asymptotique de $\hat{q}_2(.,t)$ (le seul estimateur possible dans la situation **(S.3)**) dépend d'une part du comportement de $\hat{q}_1(.,t)$ et d'autre part de comportement de $\hat{\gamma}_n(t)$. Remarquons que l'estimateur $\hat{q}_2(.,t)$ peut être utilisé dans les trois situations.

4 Une illustration sur simulation

Comme estimateur de l'indice de queue conditionnel, nous utilisons la famille d'estimateurs proposée par Gardes et Girard (2008). Elle est définie par

$$\hat{\gamma}_n(t,W) = \frac{1}{k_t}\sum_{i=1}^{k_t} i \log\left(\frac{Z_{m_t-i+1,m_t}}{Z_{m_t-i,m_t}}\right) W(i/k_t,t) \Big/ \sum_{i=1}^{k_t} W(i/k_t,t)\ , \qquad (4)$$

où $W(.,t)$ est une fonction de poids définie sur $]0,1[$ dont l'intégrale vaut 1 et $k_t = m_t\beta_{m_t}$. Soit $E = [0,1]$, on définit la fonction

$$x \in E \mapsto \gamma(x) = 1/2 - (x-1/2)^2$$

et on simule un échantillon $\{(Y_i, x_i), i = 1, ..., n\}$ de taille $n = 1000$ et de loi de Fréchet. Pour une telle loi, le quantile conditionnel est donné par,

$$q(\alpha,t) = \left\{ \log\left(\frac{1}{1-\alpha}\right)\right\}^{-\gamma(t)}.$$

Afin d'estimer $q(\alpha,t)$, on utilise $\hat{q}_2(\alpha,t)$ pour lequel on estime γ en utilisant une fonction de poids logarithmique définie par $W(s,t) = -\log(s)$. Enfin on pose $\beta_{m_t} = 0.3$ et on fixe le rayon de la boule à $h_t = 0.1$. Les résultats obtenus sont présentés sur la figure 1. Le choix des paramètres h_t et β_{m_t} est un problème délicat. On peut cependant utiliser la méthode proposée dans Gardes et Girard (2008) basée sur l'étude de la différence entre deux estimateurs différents de γ.

Bibliographie

[1] Beirlant, J. and Goegebeur, Y. (2004) Local polynomial maximum likelihood estimation for Pareto-type distributions, *Journal of Multivariate Analysis*, 89, 97–118.

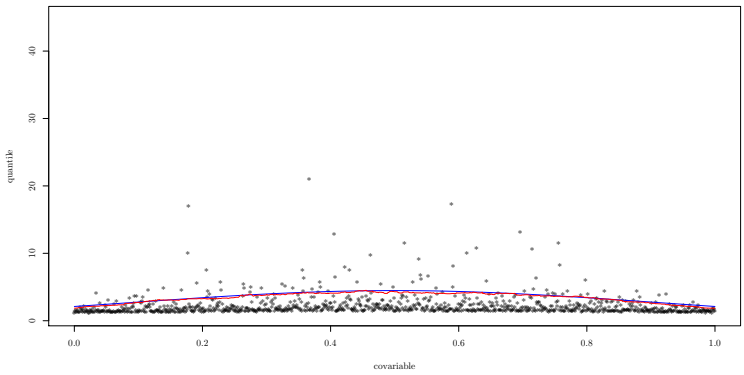
[2] Bingham, N.H., Goldie, C.M. and Teugels, J.L. (1987) *Regular variation*, Encyclopedia of Mathematics and its Applications, 27, Cambridge University Press.

[3] Ferraty, F. and Vieu, P. (2006) *Nonparametric Functional Data Analysis : Theory and Practice*, Springer Series in Statistics, Springer.

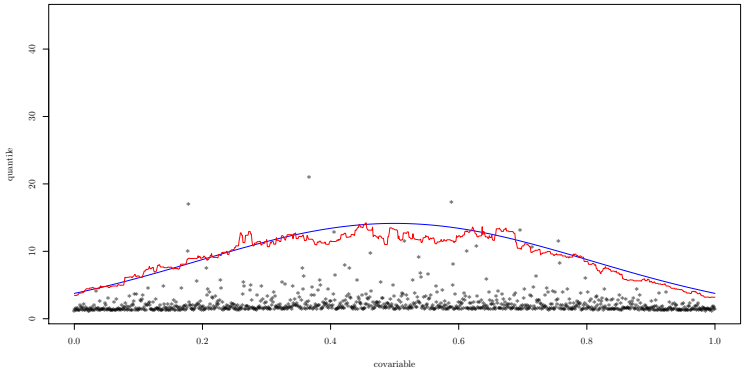
[4] Gardes, L. and Girard, S. (2008) A moving window approach for nonparametric estimation of the conditional tail index, *Journal of Multivariate Analysis*, 99, 2368–2388.

[5] Gardes, L., Girard, S. and Lekina, A. (2008) Functional nonparametric estimation of conditional extreme quantiles, `http ://hal.archives-ouvertes.fr/hal-00289996/fr/`.

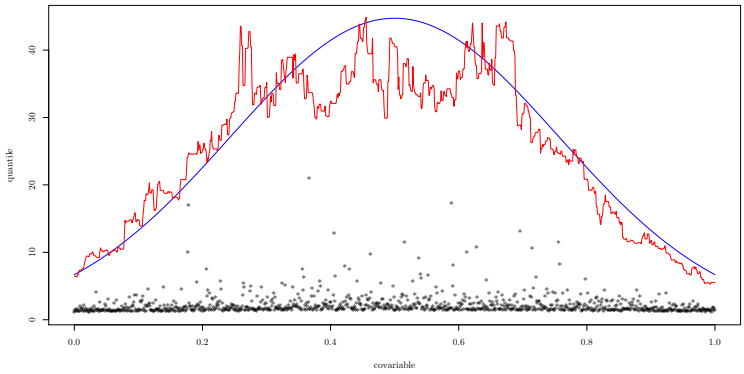
[7] Weissman, I. (1978), Estimation of parameters and large quantiles based on the k-largest observations, *Journal of the American Statistical Association*, 73, 812–815.



(S.1) avec $\alpha = 0.1$



(S.2) avec $\alpha = 0.01$



(S.3) avec $\alpha = 0.001$

FIG. 1 – Estimation de la fonction $q(\alpha, .)$ (en bleu) par $\hat{q}_2(\alpha, .)$ (en rouge).