HAL open science

Estimation de courbes de niveaux extrêmes pour des lois à queues lourdes

Abdelaati Daouia, Laurent Gardes, Stéphane Girard, Alexandre Lekina

To cite this version:

Abdelaati Daouia, Laurent Gardes, Stéphane Girard, Alexandre Lekina. Estimation de courbes de niveaux extrêmes pour des lois à queues lourdes. 42èmes Journées de Statistique, Société Française de Statistique, May 2010, Marseille, France. inria-00494684

HAL Id: inria-00494684

https://inria.hal.science/inria-00494684v1

Submitted on 24 Jun 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Estimation de courbes de niveaux extrêmes pour des lois à queues lourdes

Abdelaati Daouia[1], Laurent Gardes[2], Stéphane Girard[2] & Alexandre Lekina[2]

[1] *GREMAQ, Université de Toulouse, Aile J.J. Laffont*
21 allée de Brienne, 31000 Toulouse, France.
[2] *Équipe Mistis, INRIA Rhône-Alpes & Laboratoire Jean Kuntzmann*
Inovallée, 655, av. de l'Europe, Montbonnot, 38334 Saint-Ismier cedex, France.

Résumé

Le problème d'estimation des courbes de niveaux extrêmes est équivalent à l'étude des quantiles conditionnels quand l'ordre du quantile converges vers un. Nous montrons que sous certaines conditions, il est possible d'estimer de telles courbes au moyen d'un estimateur à noyau de la fonction de survie conditionnelle. En conséquence, ce résultat nous permet d'introduire deux versions lisses de l'estimateur de l'indice de queue conditionnel indispensable lorsque l'on veut extrapoler. Nous établissons la loi limite des estimateurs ainsi construits. Pour conclure, une illustration sur données simulées est présentée.

Mots-clés : Extrêmes, quantile conditionnel, estimateur à noyau.

Abstract

The problem of estimating extreme level curves is equivalent to studying the conditional quantiles when the order of the quantile converges to one as the sample size increases. We show that under some conditions, it is possible to estimate these curves using a kernel estimator of the conditional survival function. As a consequence, this result allows us to introduce two smooth versions of the conditional tail index estimator necessary to extrapolate. Asymptotic distributions of these estimators are established. To conclude, an illustration on simulated data is presented.

Keywords : Extreme-values, conditionnal quantile, kernel estimator.

1 Introduction

Soient $\{(X_i, Y_i), i = 1, \ldots, n\}$ des copies indépendantes du couple aléatoire $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ où Y est une variable d'intérêt associée à une covariable X. Pour tout $x \in \mathbb{R}^d$ et pour toute suite réelle $\alpha_n \to 0$, on se propose d'estimer les courbes de niveaux extrêmes définies comme les graphes de fonctions $x \in \mathbb{R}^d \mapsto q(\alpha_n|x) \in \mathbb{R}$ vérifiant

$$\mathbb{P}\left(Y > q(\alpha_n|x)|X = x\right) = \alpha_n,$$

lorsque la fonction de répartition conditionnelle de Y sachant $X = x$ est à queue lourde. De façon plus précise, ceci signifie que pour tout $y > 0$,

$$\bar{F}(y|x) \stackrel{def}{=} 1 - F(y|x) = y^{-1/\gamma(x)}\ell(y|x),$$

avec $\gamma(.)$ une fonction inconnue et positive de la covariable x que l'on appelle *indice de queue conditionnel* et $\ell(.|x)$ une fonction à variations lentes à l'infini que l'on supposera normalisée, c'est-à-dire qu'elle peut se réécrire sous la forme

$$\ell(y|x) = c(x)\exp\left(\int_1^y \frac{\varepsilon(u|x)}{u}du\right),$$

avec $c(.)$ une fonction positive et $\varepsilon(y|x) \to 0$ quand $y \to \infty$.

Le lecteur pourra se référer à Bingham, Goldie et Teugels (1987) pour une présentation plus détaillée de la théorie des fonctions à variations lentes.

2 Définition de nos estimateurs

Un estimateur naturel de la fonction $x \in \mathbb{R}^d \mapsto q(\alpha_n|x)$ appelée *quantile conditionnel* est donné par

$$\hat{q}_n(\alpha_n|x) \stackrel{def}{=} \hat{\bar{F}}_n^{\leftarrow}(\alpha_n|x) = \inf\left\{t,\ \hat{\bar{F}}_n(t|x) \leq \alpha_n\right\}. \tag{1}$$

Afin d'estimer la fonction de survie conditionnelle, on se propose d'utiliser un estimateur à noyau introduit par Collomb (1976). Il est défini par

$$\hat{\bar{F}}_n(y|x) = \frac{1}{nh_n^d}\sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right)\mathbb{1}_{\{Y_i \geq y\}} \Bigg/ \frac{1}{nh_n^d}\sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right),$$

où la fonction $K(.)$ appelée noyau est une fonction positive, bornée, intégrable et à support compact $S \subseteq \mathbb{R}^d$ et h_n est une suite non aléatoire telle que $h_n \to 0$ quand $n \to \infty$ appelée *paramètre de lissage*. Le dénominateur de la fonction de survie conditionnelle est l'estimateur à noyau classique de la densité $g(.)$ de X.

Dans cet article, nous nous intéressons à l'estimation du réel $q(\alpha_n|x)$ lorsque $\alpha_n \to 0$. Nous parlons de quantile extrême lorsque $\alpha_n \to 0$ quand $n \to \infty$. Deux situations sont alors envisagées en fonction de la vitesse de convergence de α_n vers zéro.

(a) Dans la première situation, la suite α_n *converge lentement vers zéro* en ce sens que $nh_n^d\alpha_n \to \infty$. Autrement dit, ceci revient à supposer que le quantile $q(\alpha_n|x)$ ne tend pas trop vite vers l'infini quand $n \to \infty$. Dans une telle situation, l'estimation du quantile extrême conditionnel requiert d'interpoler à l'intérieur de l'ensemble des données car il y a presque sûrement un point de l'échantillon dans la région $B(x, h_n) \times (q(\alpha_n|x), \infty)$ de $\mathbb{R}^{d+1}$ où $B(x, h_n)$ est une boule centrée en x de rayon h_n (voir lemme 1). On se propose alors d'estimer le quantile extrême conditionnel par (1).

(b) Dans la deuxième situation, on autorise la suite α_n à converger vers zéro plus vite que dans la situation **(a)**, c'est-à-dire que l'on ne suppose plus que $nh_n^d\alpha_n \to \infty$. On dit alors que α_n *converge rapidement vers zéro*. De façon équivalente, il est supposé ici que le quantile $q(\alpha_n|x)$ tend vite vers l'infini quand $n \to \infty$. Dans une telle situation, l'estimation du quantile conditionnel peut nécessiter d'extrapoler au-delà des observations. On propose donc d'adapter l'estimateur de Weissman (1978) au cas conditionnel. On estime alors $q(\alpha_n|x)$ par

$$\hat{q}_n^W(\alpha_n|x) = \hat{q}_n(\beta_n|x)(\alpha_n/\beta_n)^{-\hat{\gamma}_n(x)},$$

où β_n est telle que $nh_n^d\beta_n \to \infty$. $\hat{\gamma}_n(x)$ est un estimateur de l'indice de queue conditionnel dont nous donnons deux exemples.

3 Lois asymptotiques

Il convient tout d'abord de donner quelques conditions et résultats utiles pour établir la loi asymptotique de nos estimateurs. La démonstration des résultats introduits dans cette partie peut être consultée dans Daouia, Gardes, Girard et Lekina (2010). Dans tout ce qui suit, on désigne par $d(x, x')$ la distance entre deux points $(x, x') \in \mathbb{R}^d \times \mathbb{R}^d$ et on fait les trois hypothèses suivantes dites de régularité.

(L.1) : Il existe $c_\gamma > 0$ tel que $\left|\dfrac{1}{\gamma(x)} - \dfrac{1}{\gamma(x')}\right| \leq c_\gamma d(x, x')$.

(L.2) : Il existe $c_\ell > 0$ et $y_0 > 1$ tel que $\sup\limits_{y \geq y_0}\left|\dfrac{\log \ell(y|x)}{\log y} - \dfrac{\log \ell(y|x')}{\log y}\right| \leq c_\ell d(x, x')$.

(L.3) : Il existe $c_g > 0$ tel que $\left|g(x) - g(x')\right| \leq c_g d(x, x')$.

Dans le but de contrôler le comportement de la fonction de survie conditionnelle par rapport à sa première variable, on introduit la condition suivante sur le terme de biais $\varepsilon(.|x)$.

(F) : La fonction $|\varepsilon(.|x)|$ est asymptotiquement décroissante.

Le contrôle de ce terme est d'une importance capitale quand on cherche à établir des résultats sur la normalité asymptotique des estimateurs de l'indice de queue conditionnel.

Le lemme suivant donne une interprétation géométrique de la condition $nh_n^d\alpha_n \to \infty$ introduite précédemment.

Lemme 1 *Supposons les conditions* **(L.1)**, **(L.2)** *et* **(L.3)** *satisfaites. Considérons la région de* $\mathbb{R}^{d+1}$ *définie par* $R_n(x) = B(x, h_n) \times (q(\alpha_n|x), \infty)$ *où* $x \in \mathbb{R}^d$ *est tel que* $g(x) > 0$. *Si* $h_n \log q(\alpha_n|x) \to 0$ *quand* $n \to \infty$, *alors* $\mathbb{P}(\exists i \in \{1, \ldots, n\}, (X_i, Y_i) \in R_n(x)) \to 1$ *quand* $n \to \infty$ *si et seulement si,* $nh_n^d\alpha_n \to \infty$.

Le théorème suivant établit la normalité asymptotique de l'estimateur $\hat{q}_n(.|x)$.

Théorème 1 *Supposons les conditions* **(L.1)**, **(L.2)** *et* **(L.3)** *satisfaites. Soit* $\{\tau_j, j = 1, \ldots, J\}$ *une suite strictement positive et décroissante. On pose* $\alpha_{n,j} = \tau_j \alpha_n$. *Si* $\alpha_n \to 0$, $nh_n^d \alpha_n \to \infty$ *et* $nh_n^{d+2}\alpha_n \log^2(\alpha_n) \to 0$ *quand* $n \to \infty$, *alors, pour tout* $x \in \mathbb{R}^d$ *tel que* $g(x) > 0$, *on a*

$$\left\{\sqrt{nh_n^d \alpha_n}\left(\frac{\hat{q}_n(\alpha_{n,j}|x)}{q(\alpha_{n,j}|x)} - 1\right)\right\}_{\{j=1,\ldots,J\}} \xrightarrow{\mathcal{L}} \mathcal{N}\left(0_{\mathbb{R}^J}, \gamma^2(x)\frac{\|K\|_2^2}{g(x)}\Sigma\right),$$

où $\Sigma_{j,j'}(x) = 1/\tau_{j\wedge j'}$ *pour* $(j, j') \in \{1, \ldots, J\}^2$.

Dans la situation **(a)**, la variance asymptotique étant inversement proportionnelle à $nh_n^d \alpha_n$, l'estimation des courbes de niveaux extrêmes est d'autant plus stable que l'on s'éloigne de la frontière de l'ensemble des données. Aussi, comme cette variance est proportionnelle à $\gamma^2(x)$, ceci implique que l'estimation de $q(\alpha_n|x)$ est plus difficile pour des grandes valeurs de l'indice de queue conditionnel.

De ce théorème, on déduit deux estimateurs de l'indice de queue conditionnel. L'intérêt de construire de tels estimateurs est double. D'une part il nous permet de construire des intervalles de confiance du quantile extrême conditionnel $q(\alpha_n|x)$ et d'autre part de pouvoir extrapoler au-delà des observations. Le premier estimateur de $\gamma(x)$ que nous proposons est de type Pickands (1975) et il est donné par

$$\hat{\gamma}_n^P(x) = \frac{1}{\log 2}\log\left(\frac{\hat{q}_n(\alpha_n|x) - \hat{q}_n(2\alpha_n|x)}{\hat{q}_n(2\alpha_n|x) - \hat{q}_n(4\alpha_n|x)}\right).$$

Notre second estimateur est de type Hill (1975). Il est défini pour tout $J > 1$ par

$$\hat{\gamma}_n^H(x) = \sum_{j=1}^{J}\log\left(\hat{q}_n(\alpha_{n,j}|x)/\hat{q}_n(\alpha_{n,1}|x)\right) \Big/ \sum_{j=1}^{J}\log(\tau_1/\tau_j).$$

Corollaire 1 *Supposons la condition* **(F)** *vérifiée. Sous les hypothèses du Théorème 1 :*

(i) Si $\sqrt{nh_n^d\alpha_n}\varepsilon(q(2\alpha_n|x)|x) \to 0$ *quand* $n \to \infty$, *alors*

$$\sqrt{nh_n^d\alpha_n}\left(\hat{\gamma}_n^P(x) - \gamma(x)\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{\|K\|_2^2}{g(x)}\frac{\gamma^2(x)(2^{2\gamma(x)+1}+1)^2}{4(\log 2)^2(2^{\gamma(x)}-1)^2}\right).$$

(ii) Si $\sqrt{nh_n^d\alpha_n}\varepsilon(q(\alpha_{n,1}|x)|x) \to 0$ *quand* $n \to \infty$, *alors*

$$\sqrt{nh_n^d\alpha_n}\left(\hat{\gamma}_n^H(x) - \gamma(x)\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{V_J\|K\|_2^2}{g(x)}\gamma^2(x)\right),$$

où $V_J = \left(\sum_{j=1}^{J}\frac{2(J-j)+1}{\tau_j} - J^2\right) \Big/ \left(\sum_{j=1}^{J}\log(\tau_1/\tau_j)\right)^2$.

Par comparaison, la variance asymptotique de l'estimateur $\gamma_n^P(.)$ (resp. $\gamma_n^H(.)$) est à un facteur d'échelle $\|K\|_2^2/g(x)$ (resp. $V_J\|K\|_2^2/g(x)$) près identique à celle de l'estimateur classique de Pickands (resp. de Hill) dans le cas non conditionnel.

Théorème 2 *Supposons les conditions* **(L.1)**, **(L.2)** *et* **(L.3)** *satisfaites. Soit β_n une suite positive tendant vers zéro telle que $nh_n^d\beta_n \to \infty$ et $nh_n^{d+2}\beta_n \log^2(\beta_n) \to 0$ quand $n \to \infty$. Soit $\hat{\gamma}_n(x)$ un estimateur de l'indice de queue tel que*

$$\sqrt{nh_n^d\beta_n}\left(\hat{\gamma}_n(x) - \gamma(x)\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, v^2(x)\right) \text{ avec } v(x) > 0. \tag{2}$$

Si $\alpha_n/\beta_n \to 0$ alors, pour tout $x \in \mathbb{R}^d$, on a

$$\frac{\sqrt{nh_n^d\beta_n}}{\log\left(\beta_n/\alpha_n\right)}\left(\frac{\hat{q}_n^W\left(\alpha_n|x\right)}{\hat{q}_n\left(\alpha_n|x\right)} - 1\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, v^2(x)\right).$$

La loi asymptotique de $\hat{q}_n^W\left(.|x\right)$ dépend du comportement de $\hat{\gamma}_n(.)$. Si l'estimateur $\hat{\gamma}_n(.)$ converge moins vite que dans (2), alors la loi limite de $\hat{q}_n^W\left(.|x\right)$ peut dépendre du comportement de $\hat{q}_n\left(.|x\right)$. Pour un exemple d'une telle situation, on pourra se référer à Gardes, Girard et Lekina (2010). Aussi, remarquons que l'estimateur $\hat{q}_n^W\left(.|x\right)$ peut être utilisé dans les deux situations **(a)** et **(b)**.

4 Illustration numérique

On génère un échantillon $\{(X_i, Y_i), i = 1, ..., n\}$ de taille $n = 1000$ suivant la loi du couple $(X, Y) \in \mathbb{R} \times \mathbb{R}$ où X est une covariable de loi uniforme standard et dont le quantile conditionnel de Y sachant $X = x$ est donné par la loi de Fréchet

$$q(\alpha|x) = \{-\log(1-\alpha)\}^{-\gamma(x)}.$$

La fonction indice de queue conditionnel est définie par

$$x \in [0,1] \mapsto \gamma(x) = \frac{1}{2}\left(\frac{1}{10} + \sin(\pi x)\right)\left(\frac{11}{10} - \frac{1}{2}\exp\left(-64\left(x - 1/2\right)^2\right)\right).$$

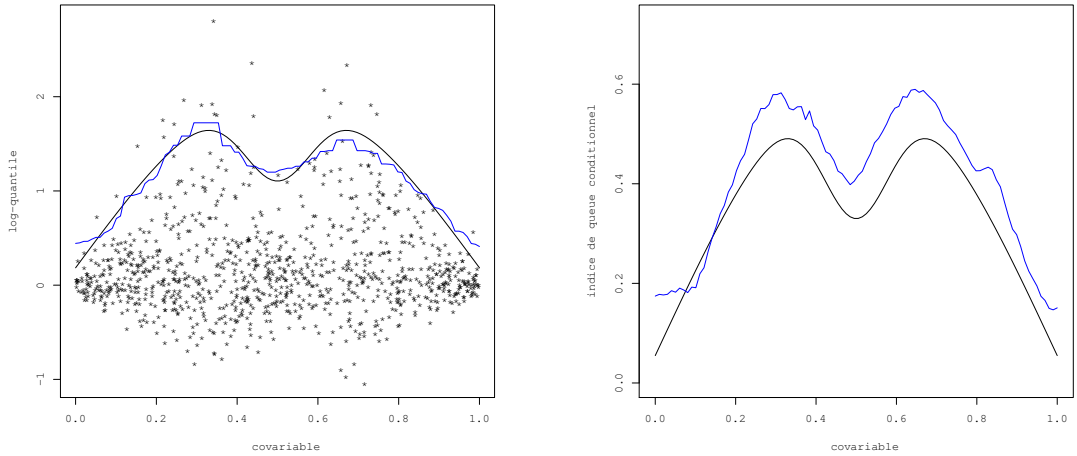
Pour cette illustration, on utilise un noyau bi-quadratique d'expression

$$K(x) = \frac{15}{16}\left(1 - x^2\right)^2 \mathbb{1}_{\{|x| \leq 1\}}.$$

On choisit le paramètre de lissage par validation croisée suivant le critère de Yao (1999)

$$h_{cv} = \arg\min_{h_n} \sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\mathbb{1}_{\{Y_i \geq Y_j\}} - \hat{\bar{F}}_{n,-i}\left(Y_j|X_i\right)\right\}^2,$$

où $\hat{\bar{F}}_{n,-i}$ est l'estimateur de $\bar{F}$ calculé à partir de l'échantillon $\{(X_k, Y_k), k = 1, \ldots, n\}$ privé de sa ième observation (X_i, Y_i). On obtient $h_{cv} = 0.164$. À la suite de quoi, on



(G) : $q\left(\alpha_n|x\right)$ en noir et $\hat{q}_n\left(\alpha_n|x\right)$ en bleu **(D)** : $\gamma(.)$ en noir et $\hat{\gamma}_n^H(.)$ en bleu

FIG. 1 – Estimation d'une courbe de niveau extrême et de son indice de queue

fixe $\alpha_n = 5\log(n)/n$ et on représente l'estimateur de quantile $\hat{q}_n\left(\alpha_n|x\right)$ correspondant (voir figure 1, graphe **(G)**) puis on pose $\tau_j = 1/j$ et on représente l'estimateur $\hat{\gamma}_n^H(.)$ de variance minimale obtenu pour $J_{opt} = 9$ (voir figure 1, graphe **(D)**).

Bibliographie

[1] N.H. Bingham, C.M. Goldie et J.L. Teugels (1987). *Regular Variation*, Cambridge University Press.

[2] G. Collomb (1976). *Estimation non paramétrique de la régression par la méthode du noyau*. PhD thesis, Université Paul Sabatier de Toulouse.

[3] A. Daouia, L. Gardes, S. Girard et A. Lekina (2010). Kernel estimators of extreme level curves. `http ://hal.inria.fr/inria-00393588/fr/`.

[4] L. Gardes, S. Girard et A. Lekina (2010). Functional nonparametric estimation of conditional extreme quantiles. *Journal of Multivariate Analysis*, 101, 419–433.

[5] B.M. Hill (1975). A simple general approach to inference about the tail of a distribution. *Annals of Statistics*, 3, 1163–1174.

[6] J. Pickands (1975). Statistical inference using extreme order statistics. *Annals of Statistics*, 3, 119–131.

[7] Q. Yao (1999). Conditional predictive regions for stochastic processes. *Technical report*, University of Kent at Canterbury.