Choix de la fenêtre pour l'estimation non-paramétrique des quantiles extrêmes conditionnels

Gilles Durrieu, Ion Grama, Quang-Khoai Pham, Jean-Marie Tricot

Choix de la fenêtre pour l'estimation non-paramétrique des quantiles extrêmes conditionnels

Gilles Durrieu, Ion Grama, Quang Khoai Pham et Jean-Marie Tricot

Laboratoire de Mathématiques de Bretagne Atlantique, Université de Bretagne Sud et UMR CNRS 6205
Campus de Tohannic, 56017 Vannes.
{gilles.durrieu, ion.grama, quang-khoai.pham,jean-marie.tricot}@univ-ubs.fr

Résumé. Soient $X_{t_1}, \ldots, X_{t_n}$ des observations indépendantes associées aux temps $0 \leq t_1 < ... < t_n \leq T_{\max}$ où X_{t_i} a la fonction de répartition F_{t_i} et F_t est la loi conditionnelle de X sachant $T = t \in [0, T_{\max}]$. Pour chaque $t \in [0, T_{\max}]$, nous proposons un estimateur adaptatif non paramétrique de quantiles extrêmes de F_t. L'idée de notre approche consiste à ajuster la queue de la distribution F_t, avec une distribution de Pareto de paramètre $\theta_{t,\tau}$ à partir d'un seuil τ. Le paramètre $\theta_{t,\tau}$ est estimé en utilisant un estimateur non paramétrique à noyau de taille de fenêtre h basé sur les observations plus grandes que τ. Sous certaines hypothèses de régularité, nous montrons que l'estimateur adaptatif proposé de $\theta_{t,\tau}$ est consistant et nous donnons sa vitesse de convergence. Nous proposons une procédure de tests séquentiels pour déterminer le seuil τ et nous estimons le paramètre h par validation croisée et par une nouvelle approche adaptative. Enfin, nous étudions les propriétés de cette procédure sur des simulations.

Mots-clés. Valeurs extrêmes, estimateur non paramétrique à noyau, simulation, estimateur adaptatif.

Abstract. We observe independent random variables $X_{t_1}, ..., X_{t_n}$ associated to a sequence of times $0 \leq t_1 < ... < t_n \leq T_{\max}$, where X_{t_i} has distribution function F_{t_i} and F_t is the conditional distribution of X given $T = t \in [0, T_{\max}]$. For each $t \in [0, T_{\max}]$, we propose a nonparametric adaptive estimator for extreme quantiles of F_t. The idea of our approach is to adjust the tail of the distribution function F_t with a Pareto distribution of parameter $\theta_{t,\tau}$ starting from a threshold τ. The parameter $\theta_{t,\tau}$ is estimated using a non parametric kernel estimator of bandwidth h based on the observations larger than τ. Under some regularity assumptions, we prove that the adaptive estimators of $\theta_{t,\tau}$ is consistent and we determine its rate of convergence. We propose a sequential testing based procedure for the automatic choice of the threshold τ and we estimate the bandwidth h by cross validation and a new adaptive approaches. Finally, we study this procedure by simulations.

Keywords. Extreme values, Non parametric kernel estimator, simulation, adaptive estimator.

1 Modèle et estimateurs

Nous considérons un couple de variables aléatoires (X, T), où X représente la variable d'intérêt et $T \in [0, T_{\max}]$ le temps. Soit $F_t(x) = P(X \leq x | T = t)$ la distribution conditionnelle de X sachant $T = t$. On suppose que F_t est définie dans l'intervalle $[x_0, \infty)$, $x_0 \geq 0$ et que F_t possède une densité f_t strictement positive. Nous observons les variables aléatoires indépendantes $X_{t_1}, ..., X_{t_n}$ associées aux temps $0 \leq t_1 < ... < t_n \leq T_{\max}$ où X_{t_i} a la distribution F_{t_i}. Nous proposons une méthode d'estimation adaptative de la queue de distribution de F_t et des quantiles d'ordre élevé. L'idée de la méthode est de déterminer de manière adaptative un seuil τ et d'ajuster sur $[\tau, +\infty[$ une distribution de Pareto définie par

$$G_{\tau,\theta}(x) = 1 - \left(\frac{x}{\tau}\right)^{-\frac{1}{\theta}} \qquad x \in [\tau, +\infty[,$$

où le paramètre $\theta > 0$ et $\tau \geq x_0$ est la valeur inconnue du seuil. Nous obtenons ainsi un modèle semi-paramétrique défini par

$$F_{t,\tau,\theta}(x) = \begin{cases} F_t(x) & \text{if } x < \tau \\ 1 - (1 - F_t(\tau))(1 - G_{\tau,\theta}(x)) & \text{if } x \geq \tau. \end{cases} \tag{1}$$

Soit $\mathcal{K}(P, Q) = \int \log \frac{dP}{dQ}\, dP$ la divergence de Kullback-Leibler entre deux mesures équivalentes P et Q. Pour chaque $t \in [0, T]$ et $\tau \geq x_0$, le minimum de la divergence de Kullback-Leibler entre F_t et le modèle $F_{t,\tau,\theta}$ est atteint pour

$$\theta_{t,\tau} = \arg\min_{\theta \in \Theta} \mathcal{K}(F_{t,\tau}, G_{\tau,\theta}) = \int_\tau^\infty \log \frac{x}{\tau} \frac{F_t(dx)}{1 - F_t(\tau)}, \tag{2}$$

où $F_{t,\tau}$ est la fonction de répartition d'excès au dessus du seuil τ :

$$F_{t,\tau}(x) = 1 - \frac{1 - F(x)}{1 - F(\tau)}, \qquad x \geq \tau.$$

Pour chaque t fixé dans $[0, T]$ et pour $\tau \geq x_0$, nous construisons un estimateur non paramétrique à noyau K de taille de fenêtre h du paramètre fonctionnel $t \to \theta_{t,\tau}$. Nous estimons dans un premier temps la fonction $\theta_{t,\tau}$ au point t en utilisant un estimateur à noyau d'une taille de fenêtre h et dans un second temps nous donnons une procédure de sélection du seuil τ. En maximisant la quasi-log vraisemblance pondérée par rapport à θ, nous obtenons l'estimateur

$$\widehat{\theta}_{t,h,\tau} = \frac{1}{\widehat{n}_{t,h,\tau}} \sum_{X_{t_i} > \tau} W_{t,h}(t_i) \log\left(\frac{X_{t_i}}{\tau}\right), \tag{3}$$

où $W_{t,h}(t_i) = K\left(\frac{t_i - t}{h}\right)$ avec K une fonction noyau et $\widehat{n}_{t,h,\tau} = \sum_{i=1}^n W_{t,h}(t_i) 1_{\{X_{t_i} > \tau\}}$. L'estimateur semi-paramétrique de la fonction de répartition F_t est donné par

$$\widehat{F}_{t,h,\tau}(x) = \begin{cases} \widehat{F}_{t,h}(x), & x \in [x_0, \tau], \\ 1 - \left(1 - \widehat{F}_{t,h}(\tau)\right)\left(\frac{x}{\tau}\right)^{-\frac{1}{\widehat{\theta}_{t,h,\tau}}}, & x > \tau, \end{cases}$$

où

$$\widehat{F}_{t,h}(x) = \frac{1}{\sum_{j=1}^{n} W_{t,h}(t_j)} \sum_{i=1}^{n} W_{t,h}(t_i)\, 1_{\{X_{t_i} \leq x\}}$$

est la fonction de répartition empirique pondérée. L'estimateur semi-paramétrique du quantile d'ordre p est donné par

$$\widehat{q}_p(t) = \begin{cases} \widehat{F}_{t,h}^{-1}(p) & \text{pour} \quad p < \widehat{p}_0, \\ \tau \left(\frac{1-\widehat{p}_0}{1-p}\right)^{\widehat{\theta}_{t,h,\tau}} & \text{sinon,} \end{cases} \tag{4}$$

avec $\widehat{p}_0 = \widehat{F}_{t,h}(\tau)$.

La principale difficulté concerne les choix du seuil τ et de la taille de la fenêtre h. Dans les paragraphes qui suivent, nous donnons des procédures pour déterminer simultanément τ et h.

2 Choix du seuil τ

L'estimateur $\widehat{\theta}_{t,h,\tau}$ est très sensible aux choix du seuil τ et de la taille de fenêtre h. La difficulté est de choisir τ assez petit de façon à ce que l'estimateur de la fonction de répartition empirique pondérée dans le modèle (1) dispose de suffisamment d'observations pour assurer un bon ajustement de la queue de la distribution F_t. Par ailleurs, τ doit être aussi choisi assez grand de façon à éviter un biais d'estimation due à un mauvais ajustement de la queue de distribution. Nous proposons d'estimer le paramètre τ en utilisant une procédure séquentielle de tests d'adéquations similaire à celle proposée par Grama et Spokoiny (2007-2008), Durrieu et al. (2012-2014). Dans un premier temps, nous testons $\mathcal{H}_0(\tau)$ l'hypothèse nulle stipulant que F_t est défini par (1) et $s_1, ..., s_m$ une suite d'instants triés par ordre décroissant de sorte que $s_1 \geq ... \geq s_m$ avec m fixé. Dans notre cas, on choisit comme suite s_k les statistiques d'ordre dans la fenêtre de largeur h autour du point t. Nous considérons une suite de tests d'adéquation en déterminant le premier instant s_k notée $\widehat{s}$ pour lequel $\mathcal{H}_0(s_k)$ est rejetée en faveur de l'hypothèse alternative $\mathcal{H}_1(\tau)$: "$F_{t,\tau}$ est la distribution de Pareto avec un point de rupture" où $F_{t,\tau}$ est la fonction de répartition d'excès de F_t au dessus du seuil τ.

Ainsi par cette procédure, nous sélectionnons le meilleur modèle en maximisant par rapport à τ la fonction de vraisemblance pénalisée donnée par :

$$\mathcal{L}_{\tau,h}\left(\tau, \widehat{\theta}_{t,h,\tau}\right) - \mathrm{Pen}_{\tau,h}\left(\tau, \widehat{\theta}_{t,h,\widehat{s}}\right) \qquad \text{où} \qquad \mathrm{Pen}_{t,h}(\tau,\theta) = \mathcal{L}_{t,h}(\tau,\theta),$$

et

$$\mathcal{L}_{t,h}(\tau,\theta) = \sum_{i=1}^{n} W_{t,h}(t_i) \log \frac{dF_{t,\tau,\theta}}{dx}(X_{t_i}).$$

Nous notons $\widehat{\tau}_{t,h}$ le seuil ainsi obtenu.

3 Estimation de la taille de la fenêtre h

Le choix du paramètre h est un point crucial. Nous proposons deux méthodes : une basée sur une approche de type validation croisée et la seconde sur une procédure adaptative.

3.1 Validation croisée

Nous considérons $\mathcal{H} = \{h_m : h_m = h_0 q^m, m = 1, \ldots, M_h\}$ avec $q > 1$, $h_0 > 0$ et M_h grand. Nous proposons la fonction de validation croisée :

$$CV(h_m, p) = \frac{1}{M_h card(T_{grid})} \sum_{h_l \in \mathcal{H}} \sum_{t_i \in T_{grid}} \psi\left(\widehat{F}^{-1}_{t_i,h_l}(p), \widehat{q}_p^{(-i)}(t_i, h_m)\right), \tag{5}$$

où T_{grid} est une suite de points d'une grille régulière sur $[0, T_{\max}]$, $\widehat{q}_p^{(-i)}(t_i, h_m)$ désigne un estimateur du quantile d'ordre p élevé au point t_i donnée par (4) calculé sur l'échantillon privé de l'observation X_{t_i} et $\psi(x, y) = |\log x - \log y|$, $x, y > 0$. L'estimateur de h notée h_{CV} s'obtient par minimisation par rapport à h_m de la fonction $CV(h_m, p)$ pour p fixé.

3.2 Méthode adaptative

Soit $h_1 < h_2 < \ldots < h_{M_h}$ une suite de tailles de fenêtre associée à la suite de voisinages $I_1 \subset I_2 \subset \ldots \subset I_{M_h}$ du temps $t \in [0, T_{\max}]$ où $I_m = [t - h_m, t + h_m] \cap [0, T_{\max}]$. Pour chaque h_m, notons par $\widehat{\tau}_{t,h_m}$ l'estimateur du seuil τ issu de la procédure présentée dans le paragraphe (2). Nous testons les hypothèses nulles $\mathcal{H}_0(\widehat{\tau}_{t,h_m}) : F_{t_i,\widehat{\tau}_{t,h_m}} = G_{\widehat{\tau}_{t,h_m},\theta}$ pour tous les $t_i \in I_m$ contre les hypothèses alternatives $\mathcal{H}_1(\widehat{\tau}_{t,h_m})$ stipulant qu'il existe un sous-intervalle $J_m \subset I_m$ tel que $F_{t_i,\widehat{\tau}_{t,h_m}} = G_{\widehat{\tau}_{t,h_m},\theta'}$ pour tous les $t_i \in J_m$ et $F_{t_i,\widehat{\tau}_{t,h_m}} = G_{\widehat{\tau}_{t,h_m},\theta''}$ pour tous les $t_i \in I_m \setminus J_m$, $\theta' \neq \theta''$. Nous déterminons le premier instant h_m notée h^* pour lequel l'hypothèse nulle $\mathcal{H}_0(\widehat{\tau}_{t,h_m})$ est rejetée. Nous choisissons enfin la taille de fenêtre h, notée $\hat{h}_{adapt}$, qui maximise la fonction de vraisemblance pénalisée par rapport à h, $0 < h < h^*$, donnée par :

$$\widetilde{T}_{n,h} = \widetilde{n}_{t,h,\widehat{\tau}_{t,h^*}} G\left(\frac{\widetilde{\theta}_{t,h,\widehat{\tau}_{t,h^*}}}{\widetilde{\theta}_{t,h^*,\widehat{\tau}_{t,h^*}}} - 1\right),$$

où $\quad G(x) = x - \log(1 + x), \qquad$ pour $x > -1$ et

$$\begin{aligned}
\widetilde{n}_{t,h,\tau} &= \sum_{t_i \in I_{m,h}} 1_{\{X_{t_i} > \tau\}}, \qquad \widetilde{\theta}_{t,h,\tau} = \frac{1}{\widetilde{n}_{t,h,\tau}} \sum_{t_i \in I_{m,h}} 1_{\{X_{t_i} > \tau\}} \log\left(\frac{X_{t_i}}{\tau}\right), \\
\widetilde{\theta}_{t,h^*,\tau} &= \frac{1}{\widetilde{n}_{t,h^*,\tau}} \sum_{t_i \in I_m} 1_{\{X_{t_i} > \tau\}} \log\left(\frac{X_{t_i}}{\tau}\right),
\end{aligned}$$

avec $I_{m,h} = [t - h, t + h] \cap [0, T_{\max}] \subset I_m$.

4 Propriétés asymptotiques

Nous notons $\theta_{t,\tau}$ le paramètre "oracle" définie par (2). Supposons que τ_n et h_n vérifient la condition suivante

$$\sum_{i=1}^{n} W_{t,h_n}(t_i)\chi^2(F_{t_i}, F_{t_i,\tau_n,\theta_{t,\tau_n}}) = O(\log n) \qquad \text{quand} \qquad n \to \infty, \tag{6}$$

où $\chi^2(P,Q) = \int \frac{dP}{dQ}dP - 1$ est la divergence de χ^2 entre deux lois équivalentes P et Q.

Théorème 4.1 *Sous la condition (6) et* $\bar{n}_{t,h_n,\tau_n} = \sum_{i=1}^{n} W_{t,h_n}(t_i)(1-F_{t_i}(\tau_n)) \to \infty$ *quand* $n \to \infty$, *nous avons*

$$\mathcal{K}\left(\widehat{\theta}_{t,h_n,\widehat{\tau}_{t,h_n}}, \theta_{t,\tau_n}\right) = O_P\left(\frac{\log n}{\bar{n}_{t,h_n,\tau_n}}\right) \qquad \textit{quand} \qquad n \to \infty.$$

On déduit du Théorème 4.1 que,

$$\mathcal{K}\left(F_{t,\tau_n}, G_{\tau_n,\widehat{\theta}_{t,h_n,\widehat{\tau}_{t,h_n}}}\right) = O_P\left(\frac{\log n}{\bar{n}_{t,h_n,\tau_n}}\right) \qquad \text{quand} \qquad n \to \infty.$$

Nous avons aussi déterminé les vitesses de convergence en considérant le modèle de Hall (Hall 1982), un modèle de mélange et le modèle de Fréchet.

5 Etude par simulation

Les propriétés de la méthode proposée sont étudiées sur des simulations en utilisant le modèle de mélange :

$$F_t(x) = C(1 - x^{-1/\theta_t}) + (1-C)\left(1 - x^{-1/\theta_t - 5}\right),\ x \geq 1, 0 \leq t \leq 1. \tag{7}$$

Nous fixons dans nos simulations un échantillon de taille $n = 50000$ avec $C = 0.75$, $m = 100$ et

$$\theta_t = 0.5 + 0.25\sin(2\pi t).$$

Nous choisissons le noyau Gaussien tronqué défini par :

$$K(x) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)1_{[-1,1]}(x),$$

Dans la Figure 1, nous observons un bon ajustement de l'estimateur $\widehat{q}_{0.99}(t)$. Des analyses similaires effectuées pour les modèles de Pareto avec point de rupture et Fréchet sur un nombre important d'échantillons donnent aussi des résultats satisfaisants. L'analyse des résultats sur 1000 simulations montre que le choix adaptatif donne des résultats sensiblement meilleurs que la méthode de validation croisée au sens du critère $ISRE$ où $ISRE$ désigne l'erreur relative intégrée.

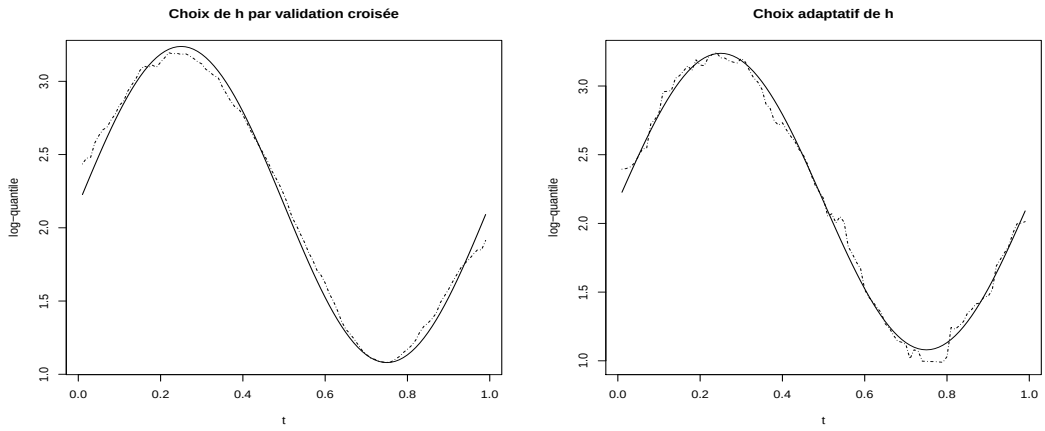

FIGURE 1 – Représentation pour un échantillon simulé des logarithmes de l'estimateur adaptatif $\widehat{q}_{0.99}(t)$ (en trait pointillé) et du 0.99-quantile théorique (en trait plein) en fonction de t. Dans la figure de gauche la taille de la fenêtre h est estimée par validation croisée, $h_{CV} = 0.076$, $ISRE = 0.0046$. Dans la figure de droite, h est choisi par la méthode adaptative, $ISRE = 0.0039$.

Bibliographie

[1] Durrieu G., Grama I., Le Tilly V., Massabuau J.C., Pham Q.K. (2012), Évènements rares sur des séries temporelles environnementales. Proc. de la société Française de Statistique, 6 pages.
[2] Durrieu G., Grama I., Pham Q.K., Tricot J.M. (2013) Estimation de quantiles extremes et probabilites rares d'un processus stochastique, Proc. de la société Française de Statistique, 6 pages.
[3] Durrieu G., Grama I., Pham Q.K., Tricot J.M. (2014) Nonparametric adaptive estimator of extreme conditional probabilities and quantiles, soumis.
[4] Grama I., Spokoiny V. (2007). Pareto approximation of the tail by local exponential modeling. Bulletin of Academy of Science of Moldova, 53(1), 1-22.
[5] Grama I., Spokoiny V. (2008) Statistics of extremes by oracle estimation, *Annals of Statistics*, 36(4), 1619-1648.
[6] Hall P. (1982) On some simple estimates of an exponent of regular variation. Journal of the Royal Statistical Society Series B, 44, 37-42.