ESTIMATION NON-PARAMETRIQUE DES QUANTILES EXTREMES CONDITIONNELS

Alexandre LEKINA

Doctorant, Equipe Mistis, INRIA Grenoble-Rhône-Alpes & LJK.
http://mistis.inrialpes.fr/people/lekina

En collaboration avec Laurent GARDES & Stéphane GIRARD

26 mai 2009

LABORATOIRE JEAN KUNTZMANN
MATHÉMATIQUES APPLIQUÉES - INFORMATIQUE

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE
INRIA
centre de recherche GRENOBLE - RHÔNE-ALPES

1. Motivation
2. Cadre d'étude statistique
3. Méthode d'estimation
4. Estimateurs et lois asymptotiques
5. Illustration par simulation

Hydrologie

En fonction de la localisation géographique x, les précipitations sont modélisées par une variable aléatoire Y. On dispose de $Y_1, ..., Y_n$ hauteurs de précipitations annuelles respectivement en n lieux $x_1, ..., x_n$ déterministes.

Hydrologie

En fonction de la localisation géographique x, les précipitations sont modélisées par une variable aléatoire Y. On dispose de $Y_1, ..., Y_n$ hauteurs de précipitations annuelles respectivement en n lieux $x_1, ..., x_n$ déterministes.

Problème :

- Calculer pour une probabilité $p < 1/n$, une hauteur de précipitation h extrême en un lieu x vérifiant $\mathbb{P}(Y > h|x) = p$.

Hydrologie

En fonction de la localisation géographique x, les précipitations sont modélisées par une variable aléatoire Y. On dispose de $Y_1, ..., Y_n$ hauteurs de précipitations annuelles respectivement en n lieux $x_1, ..., x_n$ déterministes.

Problème :

- Calculer pour une probabilité $p < 1/n$, une hauteur de précipitation h extrême en un lieu x vérifiant $\mathbb{P}(Y > h|x) = p$.

Difficultés :

- La fonction de survie conditionnelle $\mathbb{P}(Y > h|x)$ est inconnue et difficile à estimer au-delà de l'observation maximale.
- La hauteur de précipitation h est une fonction de x.

Modèle à plan fixe (design fixe)

- $Y \in \mathbb{R}$ est une v.a associée à une covariable non-aléatoire $x \in E$.
- E désigne un espace métrique muni d'une distance d.
- E peut être de dimension infinie.

Modèle à plan fixe (design fixe)

- $Y \in \mathbb{R}$ est une v.a associée à une covariable non-aléatoire $x \in E$.
- E désigne un espace métrique muni d'une distance d.
- E peut être de dimension infinie.

But

Estimer pour tout $x \in E$ et tout $\alpha \longrightarrow 0$ le réel $q(\alpha, x)$ defini par

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

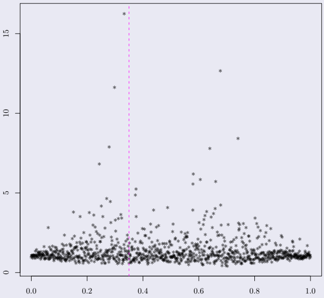
lorsque la fonction de répartition conditionnelle est dîte à queue lourde, i.e pour tout $\lambda > 0$, (se référer à Bingham, Goldie et Teugels (1987))

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

- $\gamma(x) > 0$ est une fonction inconnue de la covariable x appelée "indice des valeurs extrêmes conditionnel" (Gardes et Girard (2008)).
- $q(\alpha, x)$: "quantile extrême conditionnel" d'ordre $1 - \alpha$.

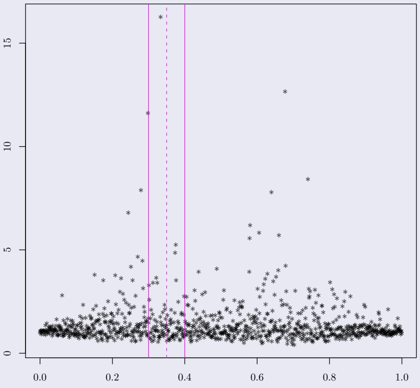
Méthode

- On dispose d'un échantillon $\{(Y_i, x_i), i = 1, \ldots, n\}$ d'observations indépendantes.
- On veut construire en un point $x \in E$ un estimateur du quantile conditionnel $q(\alpha, x)$ lorsque $\alpha \to 0$.
- Comme l'illustre ce graphique, en ordonnée on a la variable d'intérêt et en abscisse la covariable. Ici $E = [0, 1]$ et $x = 0.37$ (en trait interrompu).



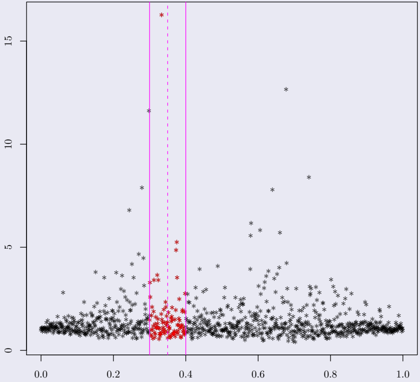
Méthode

- On utilise une méthode de fenêtres mobiles.
- $B(x, r_{n,x})$: le rayon de la boule est choisi tel que $r_{n,x} \longrightarrow 0$ quand $n \longrightarrow \infty$.



Méthode

- On ne sélectionne que les observations Y_i pour lesquelles $x_i \in B(x, r_{n,x})$.
- On note par $\{Z_i(x), i = 1, \ldots, m_{n,x}\}$ ces observations et $Z_{1,m_{n,x}}(x) \leq \ldots \leq Z_{m_{n,x},m_{n,x}}(x)$ les statistiques ordonnées correspondantes.

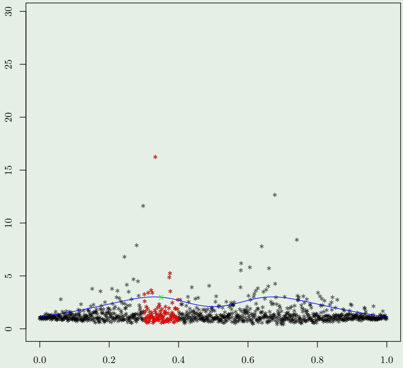


- Trois situations en fonction de la vitesse de convergence de $\alpha_{m_{n,x}}$ vers 0.

Situation (S.1)

Convergence "lente" de $\alpha_{m_{n,x}}$ vers 0, i.e $\alpha_{m_{n,x}} \longrightarrow 0$ et $m_{n,x}\alpha_{m_{n,x}} \longrightarrow \infty$.

Quantile théorique d'ordre $\alpha = 10/100 = 0.1$



Situation (S.1) Convergence "lente" : $\alpha_{m_{n,x}} \longrightarrow 0$ et $m_{n,x}\alpha_{m_{n,x}} \longrightarrow \infty$.

Estimateur

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

Loi asymptotique (Gardes, Girard & Lekina (2008))

Sous l'hypothèse que $q(., x)$ soit à variations régulières d'indice $-\gamma(x)$ (cf. équation (1)) et sous des conditions de régularités (du quantile conditionnel)

$$\left(m_{n,x}\alpha_{m_{n,x}}\right)^{1/2}\left(\frac{\hat{q}_1(\alpha_{m_{n,x}}, x)}{q(\alpha_{m_{n,x}}, x)} - 1\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \gamma^2(x)\right).$$

Situation (S.1) Convergence "lente" : $\alpha_{m_{n,x}} \to 0$ et $m_{n,x}\alpha_{m_{n,x}} \to \infty$.

Estimateur

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

Loi asymptotique (Gardes, Girard & Lekina (2008))

Sous l'hypothèse que $q(., x)$ soit à variations régulières d'indice $-\gamma(x)$ (cf. équation (1)) et sous des conditions de régularités (du quantile conditionnel)

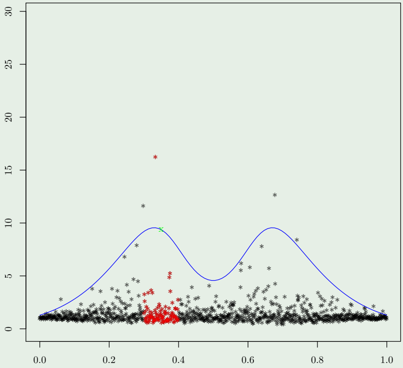
$$\left(m_{n,x}\alpha_{m_{n,x}}\right)^{1/2}\left(\frac{\hat{q}_1(\alpha_{m_{n,x}}, x)}{q(\alpha_{m_{n,x}}, x)} - 1\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \gamma^2(x)\right).$$

- La variance asymptotique est inversement proportionnelle à $\alpha_{m_{n,x}}$.
- La variance asymptotique croît avec l'indice de queue $\gamma(x)$.

Situation (S.2)

Convergence "rapide" de $\alpha_{m_{n,x}}$ vers 0, i.e $\alpha_{m_{n,x}} \to 0$ et $m_{n,x}\alpha_{m_{n,x}} \to c \geq 1$.

Quantile théorique d'ordre $\alpha = 1/100 = 0.01$



Situation **(S.2)** Convergence "rapide" : $\alpha_{m_{n,x}} \to 0$ et $m_{n,x}\alpha_{m_{n,x}} \to c \geq 1$.

Estimateur

$$\hat{q}_1(\alpha_{m_{n,x}}, x) = Z_{m_{n,x}-\lfloor m_{n,x}\alpha_{m_{n,x}}\rfloor+1, m_{n,x}}(x).$$

Loi asymptotique (Gardes, Girard & Lekina (2008))

Sous l'hypothèse que $q(.,x)$ soit à variations régulières d'indice $-\gamma(x)$ et sous des conditions de régularités (du quantile conditionnel)

$$\left(\frac{\hat{q}_1(\alpha_{m_{n,x}}, x)}{q(\alpha_{m_{n,x}}, x)} - 1\right) \xrightarrow{\mathcal{L}} \mathcal{E}\left(c, \gamma(x)\right)$$

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

Situation **(S.2)** Convergence "rapide" : $\alpha_{m_{n,x}} \to 0$ et $m_{n,x}\alpha_{m_{n,x}} \to c \geq 1$.

Estimateur

$$\hat{q}_1(\alpha_{m_{n,x}}, x) = Z_{m_{n,x}-\lfloor m_{n,x}\alpha_{m_{n,x}}\rfloor+1, m_{n,x}}(x).$$

Loi asymptotique (Gardes, Girard & Lekina (2008))

Sous l'hypothèse que $q(., x)$ soit à variations régulières d'indice $-\gamma(x)$ et sous des conditions de régularités (du quantile conditionnel)

$$\left(\frac{\hat{q}_1(\alpha_{m_{n,x}}, x)}{q(\alpha_{m_{n,x}}, x)} - 1\right) \xrightarrow{\mathcal{L}} \mathcal{E}\left(c, \gamma(x)\right)$$

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

- L'estimateur $\hat{q}_1(\alpha_{m_{n,x}}, x)$ n'est pas consistant en un certain sens i.e

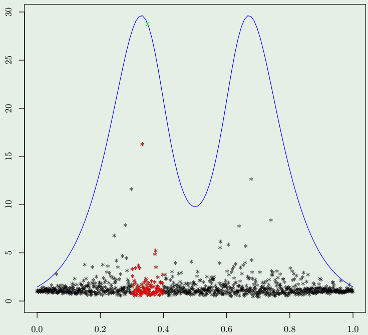
  $\dfrac{\hat{q}_1(\alpha_{m_{n,x}}, x)}{q(\alpha_{m_{n,x}}, x)}$ ne converge pas en probabilité *vers* 1 [consistance faible].

Situation (S.3)

Convergence "très rapide" de $\alpha_{m_{n,x}}$ vers 0, i.e

$$\alpha_{m_{n,x}} \longrightarrow 0 \text{ et } m_{n,x}\alpha_{m_{n,x}} \longrightarrow c \in [0, 1[.$$

Quantile théorique d'ordre $\alpha = 0.1/100 = 0.001$



Situation **(S.3)** Convergence "très rapide" de $\alpha_{m_{n,x}}$ vers 0 :

$$\alpha_{m_{n,x}} \longrightarrow 0 \text{ et } m_{n,x}\alpha_{m_{n,x}} \longrightarrow c \in [0,1[.$$

Estimateur (Weissman (1978))

$$\hat{q}_2(\alpha_{m_{n,x}}, x) = \hat{q}_1(\beta_{m_{n,x}}, x)\left(\frac{\beta_{m_{n,x}}}{\alpha_{m_{n,x}}}\right)^{\hat{\gamma}_n(x)}, \tag{3}$$

- $\beta_{m_{n,x}}$ satisfait **(S.1)**.
- $\hat{\gamma}_n(x)$ est un estimateur de l'indice des valeurs extrêmes conditionnel, (Beirlant et Goegebeur (2002) & Gardes et Girard (2008)).

Situation (S.3) Convergence "très rapide" de $\alpha_{m_{n,x}}$ vers 0 :

$$\alpha_{m_{n,x}} \longrightarrow 0 \text{ et } m_{n,x}\alpha_{m_{n,x}} \longrightarrow c \in [0,1[.$$

Estimateur (Weissman (1978))

$$\hat{q}_2(\alpha_{m_{n,x}}, x) = \hat{q}_1(\beta_{m_{n,x}}, x)\left(\frac{\beta_{m_{n,x}}}{\alpha_{m_{n,x}}}\right)^{\hat{\gamma}_n(x)}, \tag{3}$$

- $\beta_{m_{n,x}}$ satisfait **(S.1)**.
- $\hat{\gamma}_n(x)$ est un estimateur de l'indice des valeurs extrêmes conditionnel, (Beirlant et Goegebeur (2002) & Gardes et Girard (2008)).

- L'estimateur $\hat{q}_2(.,x)$ dépend non seulement de la statistique d'ordre $\hat{q}_1(.,x)$ mais aussi de l'estimateur d'indice de queue $\hat{\gamma}_n(x)$.

Situation (S.3) Convergence "très rapide" de $\alpha_{m_{n,x}}$ vers 0 :

$$\alpha_{m_{n,x}} \longrightarrow 0 \text{ et } m_{n,x}\alpha_{m_{n,x}} \longrightarrow c \in [0, 1[.$$

Loi asymptotique (Gardes, Girard & Lekina (2008))

Si la suite $\beta_{m_{n,x}}$ satisfait **(S.1)** et s'il existe une suite positive $v_n(x)$ et une loi $\mathcal{D}$ telle que

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

alors, deux situations se présentent :

(i) Soit la loi asymptotique découle de $\hat{q}_1(\beta_{m_{n,x}}, x)$ et nous avons

$$\left(m_{n,x}\alpha_{m_{n,x}}\right)^{1/2}\left(\frac{\hat{q}_2(\alpha_{m_{n,x}}, x)}{q(\alpha_{m_{n,x}}, x)} - 1\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \gamma^2(x)\right).$$

(ii) Sinon, elle provient de $\hat{\gamma}_n(x)$ et nous avons

$$\frac{v_n(x)}{\log\left(\beta_{m_{n,x}}/\alpha_{m_{n,x}}\right)}\left(\frac{\hat{q}_2(\alpha_{m_{n,x}}, x)}{q(\alpha_{m_{n,x}}, x)} - 1\right) \xrightarrow{\mathcal{L}} \mathcal{D}.$$

Simulations

- La famille d'estimateurs proposée par Gardes et Girard (2008)

$$\hat{\gamma}_n(x, W) = \sum_{i=1}^{k_{n,x}} i \log\left(\frac{Z_{m_{n,x}-i+1,m_{n,x}}}{Z_{m_{n,x}-i,m_{n,x}}}\right) W(i/k_{n,x}, x) \Bigg/ \sum_{i=1}^{k_{n,x}} W(i/k_{n,x}, x) \ ,$$

  - $W(.,x)$ est une fonction de poids définie sur $]0,1[$ dont l'intégrale vaut 1.
  - $k_{n,x} = m_{n,x}\beta_{m_{n,x}}$.

Simulations

- La famille d'estimateurs proposée par Gardes et Girard (2008)

$$\hat{\gamma}_n(x, W) = \sum_{i=1}^{k_{n,x}} i \log\left(\frac{Z_{m_{n,x}-i+1, m_{n,x}}}{Z_{m_{n,x}-i, m_{n,x}}}\right) W(i/k_{n,x}, x) \Bigg/ \sum_{i=1}^{k_{n,x}} W(i/k_{n,x}, x) \ ,$$

  - $W(., x)$ est une fonction de poids définie sur $]0, 1[$ dont l'intégrale vaut 1.
  - $k_{n,x} = m_{n,x} \beta_{m_{n,x}}$.

Remarques

- Le choix des paramètres $r_{n,x}$ et $\beta_{m_{n,x}}$ est un problème délicat.

Simulations

- La famille d'estimateurs proposée par Gardes et Girard (2008)

$$\hat{\gamma}_n(x, W) = \sum_{i=1}^{k_{n,x}} i \log\left(\frac{Z_{m_{n,x}-i+1,m_{n,x}}}{Z_{m_{n,x}-i,m_{n,x}}}\right) W(i/k_{n,x}, x) \Bigg/ \sum_{i=1}^{k_{n,x}} W(i/k_{n,x}, x) \ ,$$

  - $W(.,x)$ est une fonction de poids définie sur $]0,1[$ dont l'intégrale vaut 1.
  - $k_{n,x} = m_{n,x}\beta_{m_{n,x}}$.

Remarques

- Le choix des paramètres $r_{n,x}$ et $\beta_{m_{n,x}}$ est un problème délicat.
- Si $W(.,x) = 1$, alors

$$\hat{\gamma}_n(x) := \hat{\gamma}_n(x, 1) = \frac{1}{k_{n,x}} \sum_{i=1}^{k_{n,x}} i \log\left(Z_{m_{n,x}-i+1,m_{n,x}} / Z_{m_{n,x}-i,m_{n,x}}\right),$$

  c'est l'estimateur de Hill (1975) adapté au cas conditionnel.

Simulations

- La famille d'estimateurs proposée par Gardes et Girard (2008)

$$\hat{\gamma}_n(x, W) = \sum_{i=1}^{k_{n,x}} i \log\left(\frac{Z_{m_{n,x}-i+1,m_{n,x}}}{Z_{m_{n,x}-i,m_{n,x}}}\right) W(i/k_{n,x}, x) \Bigg/ \sum_{i=1}^{k_{n,x}} W(i/k_{n,x}, x) ,$$

  - $W(., x)$ est une fonction de poids définie sur $]0, 1[$ dont l'intégrale vaut 1.
  - $k_{n,x} = m_{n,x}\beta_{m_{n,x}}$.

Remarques

- Le choix des paramètres $r_{n,x}$ et $\beta_{m_{n,x}}$ est un problème délicat.
- Si $W(., x) = 1$, alors

$$\hat{\gamma}_n(x) := \hat{\gamma}_n(x, 1) = \frac{1}{k_{n,x}} \sum_{i=1}^{k_{n,x}} i \log\left(Z_{m_{n,x}-i+1,m_{n,x}} / Z_{m_{n,x}-i,m_{n,x}}\right),$$

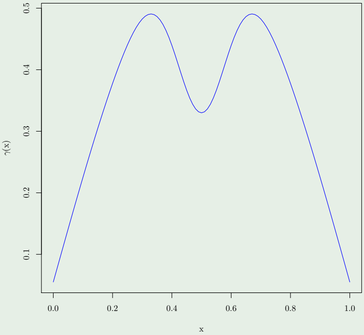
  c'est l'estimateur de Hill (1975) adapté au cas conditionnel.
- En posant $W(s, x) = -\log(s)$, on retrouve l'estimateur de Zipf (Schultze et Steinebach (1996) & Kratz et Resnick (1996)) adapté au cas conditionnel.

Simulations

- Soit $E = [0, 1]$, on définit la fonction

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

Allure de la fonction $\gamma(x)$



Simulations

- On simule un échantillon $\{(Y_i, x_i), i = 1, ..., n\}$ de taille $n = 1000$ dont le quantile conditionnel est défini par

$$q(\alpha, x) = \left\{ \log\left(\frac{1}{1-\alpha}\right)\right\}^{-\gamma(x)} \quad \text{(Loi de Fréchet).}$$

Simulations

- On simule un échantillon $\{(Y_i, x_i), i = 1, ..., n\}$ de taille $n = 1000$ dont le quantile conditionnel est défini par

$$q(\alpha, x) = \left\{ \log\left(\frac{1}{1-\alpha}\right) \right\}^{-\gamma(x)} \quad \text{(Loi de Fréchet).}$$

- On choisit arbitrairement les paramètres $\beta_{m_{n,x}}$ et $r_{n,x}$:

$$\beta_{m_{n,x}} = 0.25 \text{ et } r_{n,x} = 0.1.$$

Simulations

- On simule un échantillon $\{(Y_i, x_i), i = 1, \ldots, n\}$ de taille $n = 1000$ dont le quantile conditionnel est défini par

$$q(\alpha, x) = \left\{ \log\left(\frac{1}{1-\alpha}\right) \right\}^{-\gamma(x)} \quad \text{(Loi de Fréchet).}$$

- On choisit arbitrairement les paramètres $\beta_{m_{n,x}}$ et $r_{n,x}$:

$$\beta_{m_{n,x}} = 0.25 \text{ et } r_{n,x} = 0.1.$$

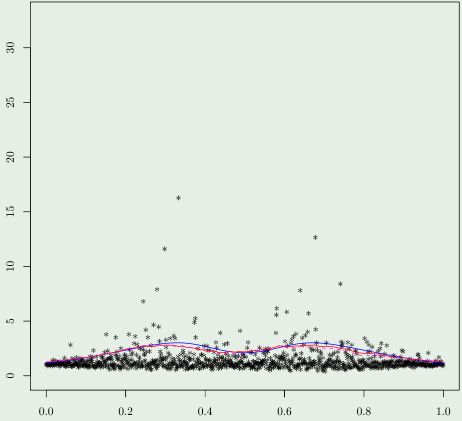
- On utilise l'estimateur

$$\hat{q}_2(\alpha_{m_{n,x}}, x) = \hat{q}_1(\beta_{m_{n,x}}, x) \left(\frac{\beta_{m_{n,x}}}{\alpha_{m_{n,x}}}\right)^{\hat{\gamma}_n(x)}.$$

- $W(s,x) = -\log(s)$ (en magenta) et $W(s,x) = 1$ (en rouge).

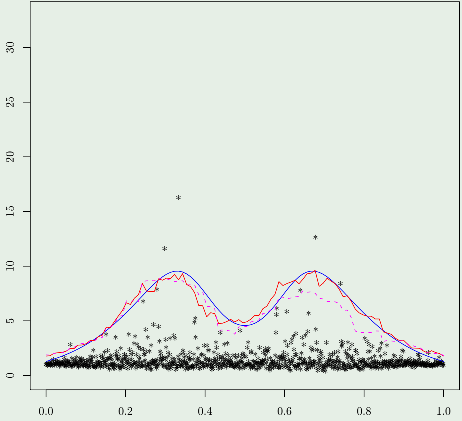
(S.1) Estimation du quantile théorique d'ordre $\alpha = 10/100 = 0.1$ avec $\hat{q}_2(\alpha, x)$

— $\hat{q}_2(\alpha, .)$ avec $W(s, x) = -\log(s)$ et — $\hat{q}_2(\alpha, .)$ avec $W(s, x) = 1$.



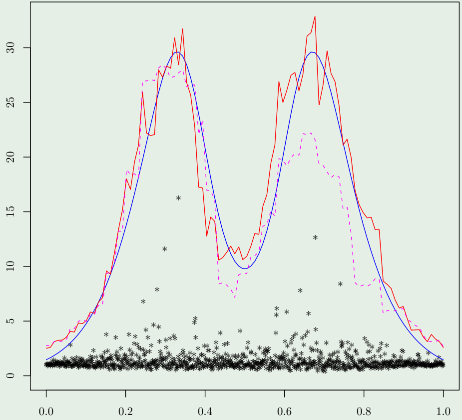
(S.2) Estimation du quantile théorique d'ordre $\alpha = 1/100 = 0.01$ avec $\hat{q}_2(\alpha, x)$

— $\hat{q}_2(\alpha, .)$ avec $W(s, x) = -\log(s)$ et — $\hat{q}_2(\alpha, .)$ avec $W(s, x) = 1$.



(S.3) Estimation du quantile théorique d'ordre $\alpha = 0.001$ avec $\hat{q}_2(\alpha, x)$

— $\hat{q}_2(\alpha, .)$ avec $W(s,x) = -\log(s)$ et — $\hat{q}_2(\alpha, .)$ avec $W(s,x) = 1$.



Bibliographie

[1] Beirlant, J. et 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. et Teugels, J.L. (1987) *Regular variation*, Encyclopedia of Mathematics and its Applications, 27, Cambridge University Press.

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

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

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

[6] Kratz, M. et Resnick, S. (1996) The QQ-estimator and heavy tails. *Stochastic Models*, 12,699–724.

[7] Schultze, J. et Steinebach, J. (1996) On least squares estimates of an exponential tail coefficient. *Statistics and Decisions*, 14, 353–372.

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

♣ QUESTIONS ...