Estimation non-paramétrique de courbes de niveaux extrêmes

Stéphane Girard

INRIA Rhône-Alpes, équipe-projet MISTIS
http://mistis.inrialpes.fr/people/girard/

Avril 2010

en collaboration avec Abdelaati Daouia, Laurent Gardes et d'après le travail de thèse d'Alexandre Lekina

Position du problème

- Statistique des valeurs extrêmes : estimation des quantiles extrêmes associés à une v.a. Y de $\mathbb{R}$ définis par

$$\mathbb{P}(Y > q(\alpha)) = \alpha,$$

quand $\alpha \to 0$.

- Statistique fonctionnelle : une covariable $X \in \mathbb{R}^p$ est mesurée avec Y et on cherche à estimer des quantiles conditionnels définis par

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

quand $\alpha \in]0, 1[$ est fixé.

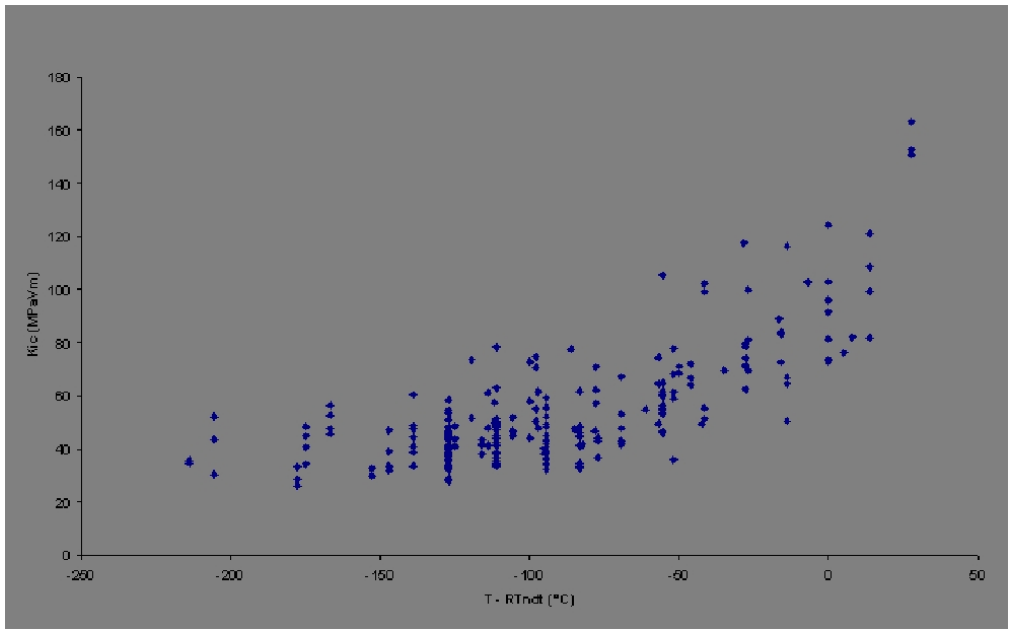
- Statistique des valeurs extrêmes "conditionnelles" : on s'intéresse aux quantiles extrêmes conditionnels définis par

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

quand $\alpha \to 0$.

Illustration : covariable unidimensionnelle ($p = 1$)

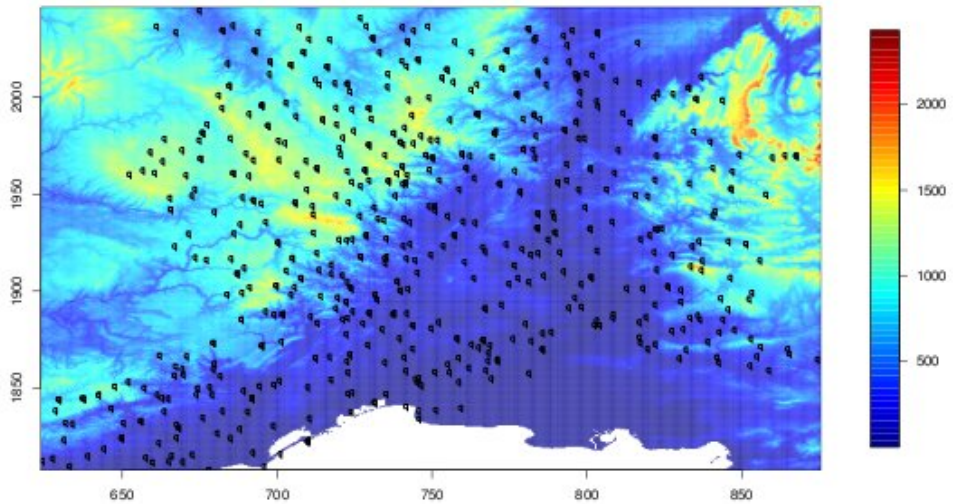
Données fournies par le Laboratoire de Conduite et Fiabilité des Réacteurs (LCFR) du CEA Cadarache.



Y est la ténacité de la cuve (verticalement), X est la température (horizontalement).

Illustration : covariable tridimensionnelle ($p = 3$)

Données fournies par le Laboratoire des Transferts en Hydrologie et Environnement (LTHE) de Grenoble dans le cadre d'une ANR.
$X = \{\text{longitude, lattitude, altitude}\}$, Y : hauteur de pluie.



Objectif : Carte des niveaux de retour moyen (en *mm*) des pluies horaires sur une période de 10 ans dans la région Cévennes-Vivarais (Gardes & Girard, 2010)

Deux problèmes "duaux"

Partant d'un échantillon (X_i, Y_i), $i = 1, \ldots, n$ de couples i.i.d.

- Estimer les quantiles extrêmes conditionnels définis par

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

quand $\alpha_n \longrightarrow 0$ lorsque $n \longrightarrow \infty$.

- Estimer les petites probabilités conditionnelles définies par

$$\bar{F}(y_n|x) \stackrel{def}{=} \mathbb{P}(Y > y_n|X = x)$$

quand $y_n \longrightarrow \infty$ lorsque $n \longrightarrow \infty$.

Principe

On utilise l'estimateur à noyau de la fonction de survie conditionnelle

$$\hat{\bar{F}}_n(y|x) = \sum_{i=1}^{n} K_h(x - X_i)\mathbb{I}\{Y_i > y\} \Big/ \sum_{i=1}^{n} K_h(x - X_i),$$

avec

- $\mathbb{I}\{.\}$ la fonction indicatrice,
- $h = h_n$ une suite déterministe tq $h \longrightarrow 0$ quand $n \longrightarrow \infty$,
- $K_h(t) = K(t/h)/h^p$ où K est une densité bornée, définie sur $\mathbb{R}^p$ et à support inclu dans la boule unité.

Hypothèse de domaine d'attraction

On suppose que la loi conditionnelle de $Y|X=x$ appartient au domaine d'attraction de Fréchet *i.e.*

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

- $\gamma(.)$ une fonction positive de la covariable x appelée indice de queue conditionnel,
- $\ell(.|x)$ une fonction à variations lentes (à x fixé) *i.e.* $\forall \lambda > 0$,

$$\lim_{y\to\infty} \frac{\ell(\lambda y|x)}{\ell(y|x)} = 1.$$

De façon équivalente, $\bar{F}(.|x)$ est à variations régulières d'indice $-1/\gamma(x)$ *i.e.* $\forall \lambda > 0$

$$\lim_{y\to\infty} \frac{\bar{F}(\lambda y|x)}{\bar{F}(y|x)} = \lambda^{-1/\gamma(x)}.$$

Hypothèses de régularité

- Notations : Soit g la densité de X. Pour tout $(x, x') \in \mathbb{R}^p \times \mathbb{R}^p$, on note $d(x, x')$ la distance entre x et x'.
- Conditions de Lipschitz : Il existe des constantes positives y_0, c_ℓ, c_γ et c_g telles que

$$\begin{aligned} \left| \frac{1}{\gamma(x)} - \frac{1}{\gamma(x')} \right| &\leq c_\gamma d(x, x'), \\ \left| g(x) - g(x') \right| &\leq c_g d(x, x'), \\ \sup_{y > y_0} \left| \frac{\log \ell(y, x)}{\log y} - \frac{\log \ell(y, x')}{\log y} \right| &\leq c_\ell d(x, x'). \end{aligned}$$

Normalité asymptotique

Théorème

Si de plus,

- $y_n \to \infty$ tq $nh^p \bar{F}(y_n|x) \to \infty$ et $nh^{p+2} \log^2(y_n) \bar{F}(y_n|x) \to 0$,
- $0 < a_1 < a_2 < \cdots < a_J$, $J \in \mathbb{N}^*$,

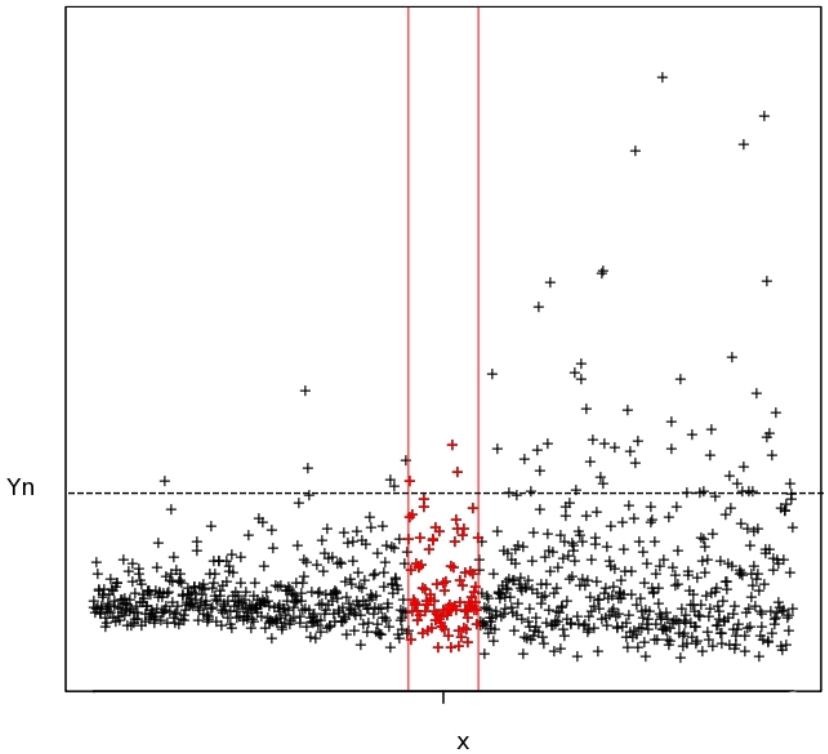
alors pour tout $x \in \mathbb{R}^p$ tel que $g(x) > 0$,

$$\left\{ \sqrt{nh^p \bar{F}(y_n|x)} \left(\frac{\hat{\bar{F}}_n(a_j y_n|x)}{\bar{F}(a_j y_n|x)} - 1 \right) \right\}_{j=1,\ldots,J}$$

est asymptotiquement gaussien centré de matrice de covariance $\frac{\|K\|_2^2}{g(x)} C(x)$ où $C_{j,j'}(x) = a_{j \wedge j'}^{1/\gamma(x)}$ pour tout $(j, j') \in \{1, \ldots, J\}^2$.

Condition $nh^p\bar{F}(y_n|x) \longrightarrow \infty$

CNS pour qu'il y ait presque sûrement au moins un point dans la région $B(x,h) \times [y_n, +\infty[$ de $\mathbb{R}^{p+1}$.



Condition $nh^{p+2}\log^2(y_n)\bar{F}(y_n|x) \longrightarrow 0$

- Condition pour que le carré du biais, de l'ordre de

$$(h\log y_n)^2,$$

soit négligeable devant la variance, de l'ordre de

$$\frac{1}{nh^p\bar{F}(y_n|x)}.$$

- Si y_n est borné, alors on retrouve la condition de normalité asymptotique classique : $nh^{p+2} \longrightarrow 0$.

Principe

On utilise l'inverse généralisé de l'estimateur de la fonction de survie conditionnelle

$$\hat{q}_n(\alpha_n|x) = \hat{\bar{F}}_n^{\leftarrow}(\alpha_n|x) = \inf\{y,\ \hat{\bar{F}}_n(y|x) \leq \alpha_n\},$$

avec $(\alpha_n) \subset]0,1[$.
Lorsque $\alpha_n = \alpha$ fixé, la normalité asymptotique de $\hat{q}_n(\alpha|x)$ est établie par exemple par (Samanta, 1989) ou (Berlinet *et al.*, 2001).

Hypothèse sur la fonction à variations lentes

- Représentation de Karamata : toute fonction à variations lentes peut s'écrire (Bingham *et al.*, 1987, Théorème 1.3.1)

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

où $c(y|x) \to c(x)$ et $\varepsilon(y|x) \to 0$ quand $y \to \infty$.

- Ici, on suppose que $\ell(.|x)$ est normalisée c'est à dire que $c(y|x) = c(x)$.

Normalité asymptotique

Théorème

Si de plus,

- $\alpha_n \to 0$ tq $nh^p\alpha_n \to \infty$ et $nh^{p+2}\log^2(\alpha_n)\alpha_n \to 0$,
- $\tau_1 > \tau_2 > \cdots > \tau_J > 0$, $J \in \mathbb{N}^*$,

alors pour tout $x \in \mathbb{R}^p$ tel que $g(x) > 0$,

$$\left\{ \sqrt{nh^p\alpha_n} \left(\frac{\hat{q}_n(\tau_j\alpha_n|x)}{q(\tau_j\alpha_n|x)} - 1 \right) \right\}_{j=1,\ldots,J}$$

est asymptotiquement gaussien centré de matrice de covariance $\|K\|_2^2 \frac{\gamma^2(x)}{g(x)}\Sigma$ où $\Sigma_{j,j'} = 1/\tau_{j\wedge j'}$ pour tout $(j,j') \in \{1,\ldots,J\}^2$.

Remarques sur la variance asymptotique

De l'ordre de

$$\|K\|_2^2 \frac{\gamma^2(x)}{g(x)} \frac{1}{nh^p\alpha_n}.$$

- Facteur supplémentaire $1/\alpha_n$ par rapport au cas $\alpha_n = \alpha$ fixé (Berlinet *et al.*, 2001, Théorème 6.4).
- Rôle de l'indice de queue conditionnel. Un modèle équivalent : $\tilde{\gamma}(x) = 1$ et densité proportionnelle à $\tilde{g}(x) = g(x)/\gamma^2(x)$.
- On peut choisir $h = \eta_n(n\alpha_n \log^2(\alpha_n))^{-1/(p+2)}$ où $\eta_n \to 0$. On obtient une variance asymptotique proportionnelle à

$$\eta_n^{-p} \left(\frac{n\alpha_n}{\log^p(\alpha_n)} \right)^{-\frac{2}{p+2}}.$$

Pour $p = 0$, variance des estimateurs des quantiles extrêmes classiques.

Remarque sur l'ordre des quantiles extrêmes

Les deux conditions $nh^p\alpha_n \to \infty$ et $nh^{p+2}\log^2(\alpha_n)\alpha_n \to 0$ entrainent

$$\frac{n\alpha_n}{\log^p(1/\alpha_n)} \to \infty,$$

qui implique

$$\alpha_n > \frac{\log^p(n)}{n}.$$

On ne peut pas estimer des quantiles "très extrêmes".

Estimation de l'indice de queue conditionnel

1) Estimateur de Pickands à noyau : adaptation de l'estimateur de Pickands (Pickands, 1975) au cas conditionnel.

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

Hypothèse sur la fonction à variations lentes

Rappel : représentation de Karamata quand $\ell(.|x)$ est normalisée :

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

On suppose de plus que $|\varepsilon(.|x)|$ est décroissante à l'infini.

$$\frac{\ell(\lambda y|x)}{\ell(y|x)} - 1 \sim \log\left(\frac{\ell(\lambda y|x)}{\ell(y|x)} \right) = \int_y^{\lambda y} \frac{\varepsilon(u|x)}{u} du,$$

lorsque $y \to \infty$ et par conséquent :

$$\left| \frac{\ell(\lambda y|x)}{\ell(y|x)} - 1 \right| \leq (1 + o(1))|\varepsilon(y|x)| \log(\lambda).$$

La fonction $\varepsilon(.|x)$ contrôle le biais des estimateurs en théorie des valeurs extrêmes.

Normalité asymptotique

Corollaire

Si de plus $\alpha_n \to 0$ tq $nh^p\alpha_n \to \infty$, $nh^{p+2}\log^2(\alpha_n)\alpha_n \to 0$ et $nh^p\alpha_n\varepsilon^2(q(2\alpha_n|x)|x) \to 0$, alors pour tout $x \in \mathbb{R}^p$ tel que $g(x) > 0$,

$$\sqrt{nh^p\alpha_n}(\hat{\gamma}_n^{\mathrm{P}}(x) - \gamma(x))$$

est asymptotiquement gaussien centré de variance

$$\frac{\|K\|_2^2}{g(x)} \frac{\gamma^2(x)(2^{2\gamma(x)+1}+1)^2}{4(\log 2)^2(2^{\gamma(x)}-1)^2}.$$

.

Estimation de l'indice de queue conditionnel

2) Estimateur de Hill à noyau : adaptation de l'estimateur de Hill (Hill, 1975) au cas conditionnel.

$$\hat{\gamma}_n^{\mathrm{H}}(x) = \sum_{j=1}^{J} [\log \hat{q}_n(\tau_j \alpha_n | x) - \log \hat{q}_n(\alpha_n | x)] \Big/ \sum_{j=1}^{J} \log(1/\tau_j) \ ,$$

avec $1 = \tau_1 > \tau_2 > \cdots > \tau_J > 0$.

Normalité asymptotique

Corollaire

Si de plus $\alpha_n \to 0$ tq $nh^p\alpha_n \to \infty$, $nh^{p+2}\log^2(\alpha_n)\alpha_n \to 0$ et $nh^p\alpha_n\varepsilon^2(q(\alpha_n|x)|x) \to 0$, alors pour tout $x \in \mathbb{R}^p$ tel que $g(x) > 0$,

$$\sqrt{nh^p\alpha_n}(\hat{\gamma}_n^{\mathrm{H}}(x) - \gamma(x))$$

est asymp. gaussien centré de variance $\|K\|_2^2\gamma^2(x)V_J/g(x)$ avec

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

.

Exemple : Si $\tau_j = 1/j$ alors $V_J = J(J-1)(2J-1)/(6\log^2(J!))$. est minimum pour $J = 9$ et $V_9 \simeq 1.25$.

Application à l'estimation des quantiles extrêmes conditionnels

Estimateur de Weissman à noyau : adaptation de l'estimateur de Weissman (Weissman, 1978) au cas conditionnel.

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

avec

- $\hat{q}_n(\alpha_n|x)$ l'estimateur à noyau précédent,
- $\hat{\gamma}_n(x)$ un estimateur de l'indice de queue conditionnel.

Le facteur $(\alpha_n/\beta_n)^{\hat{\gamma}_n(x)}$ permet d'extrapoler et d'estimer des quantiles extrêmes conditionnels d'ordres arbitrairement petits.

Normalité asymptotique

Théorème

Si de plus,

- $\alpha_n \to 0$ tq $nh^p\alpha_n \to \infty$ et $nh^{p+2}\log^2(\alpha_n)\alpha_n \to 0$,
- $\beta_n/\alpha_n \to 0$
- $\sqrt{nh^p\alpha_n}(\hat{\gamma}_n(x) - \gamma(x)) \xrightarrow{d} \mathcal{N}(0, v^2(x))$,

alors pour tout $x \in \mathbb{R}^p$

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

Illustration sur simulations

- Echantillon $\{(Y_i, X_i), i = 1, \ldots, n\}$ de taille $n = 300$ simulé suivant le modèle $X \sim \mathcal{U}[0,1]$ et Y sachant $X = x$ est distribué selon une loi de Fréchet *i.e.* $\bar{F}(y|x) = \exp(-y^{-1/\gamma(x)})$, avec

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

- **But :** estimation du quantile extrême conditionnel $q(\alpha_n|x) = (-\log \alpha_n)^{-\gamma(x)}$ d'ordre $\alpha_n = 5\log(n)/n$.
- Estimateur à noyau bi-quadratique :

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

Choix de la fenêtre de lissage

- Approche préconisée : validation croisée (Yao, 1999)

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

où $\hat{\bar{F}}_{n,-i}$ est l'estimateur à noyau (dépendant de h) calculé sur l'échantillon $\{(X_\ell, Y_\ell),\ 1 \leq \ell \leq n,\ \ell \neq i\}$.

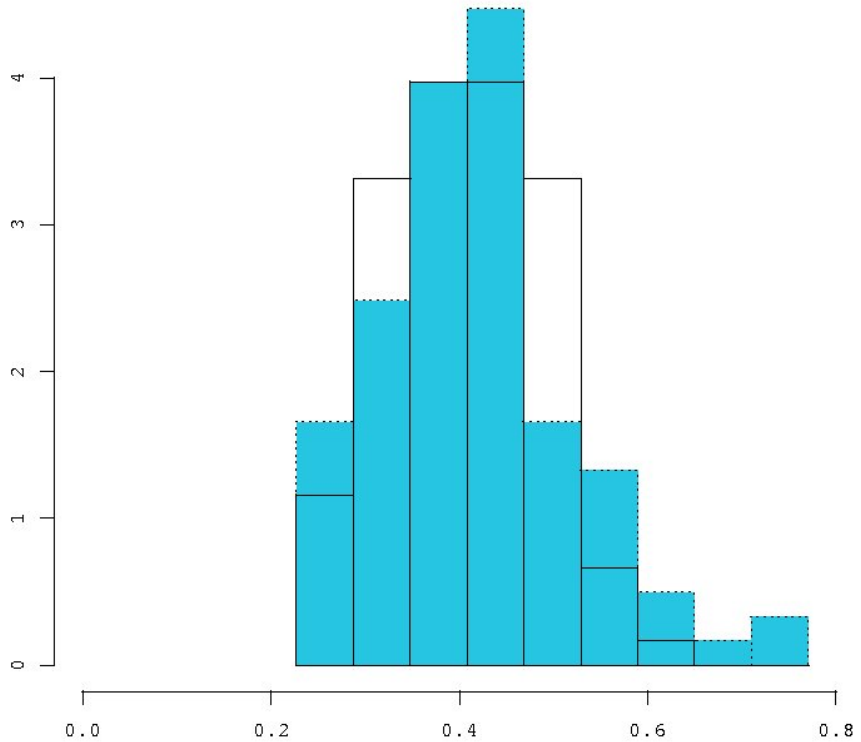
- Approche de référence : oracle. Minimisation de la distance Δ entre quantile estimé et quantile théorique :

$$h_{oracle} = \arg\min_h \Delta(q(\alpha_n|.), \hat{q}_n(\alpha_n|.)),$$

avec

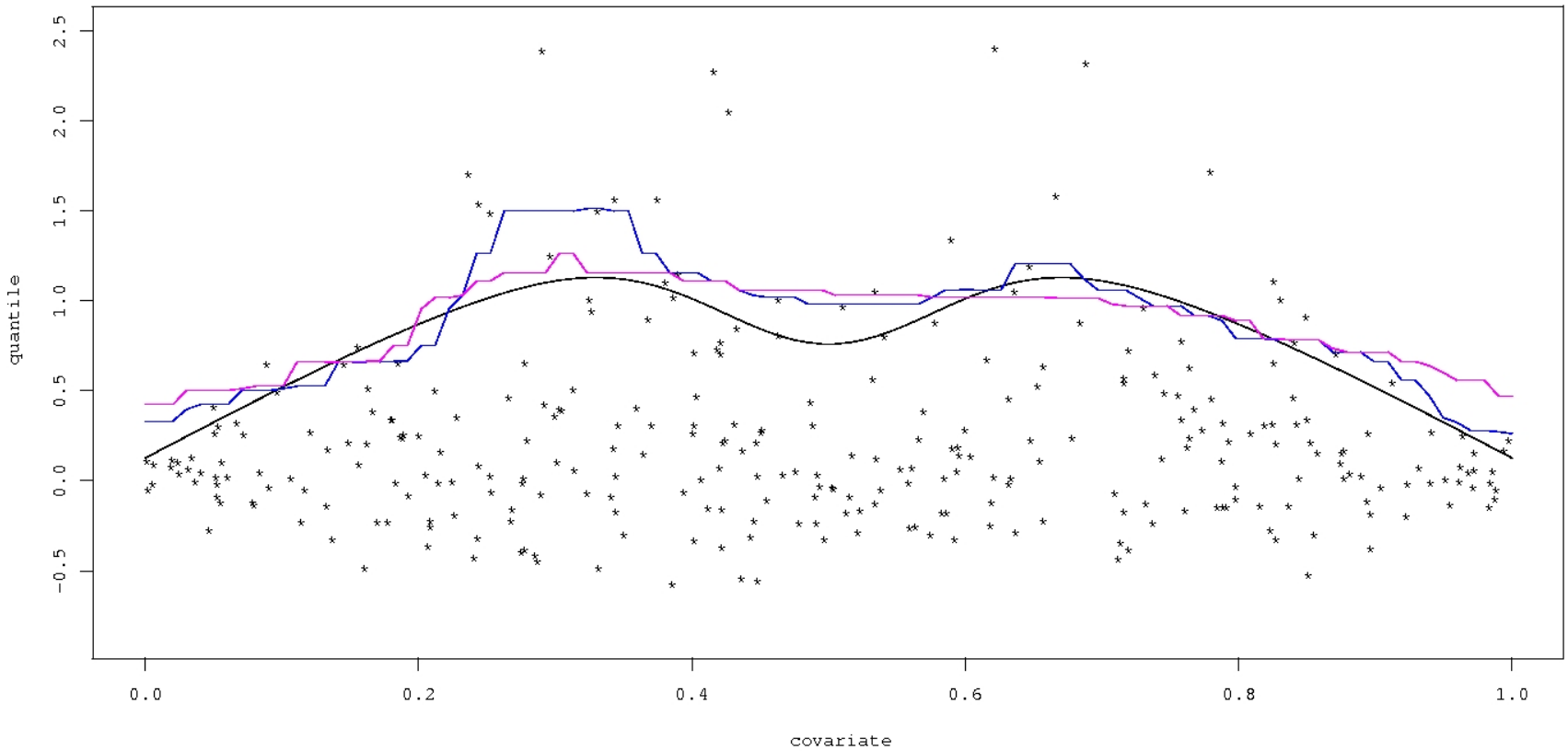
$$\Delta(q(\alpha_n|.), \hat{q}_n(\alpha_n|.)) = \left\{ \frac{1}{L} \sum_{\ell=1}^{L} (\hat{q}_n(\alpha_n|t_\ell) - q(\alpha_n|t_\ell))^2 \right\}^{1/2}.$$

Histogramme des erreurs (calculé sur 100 réplications)



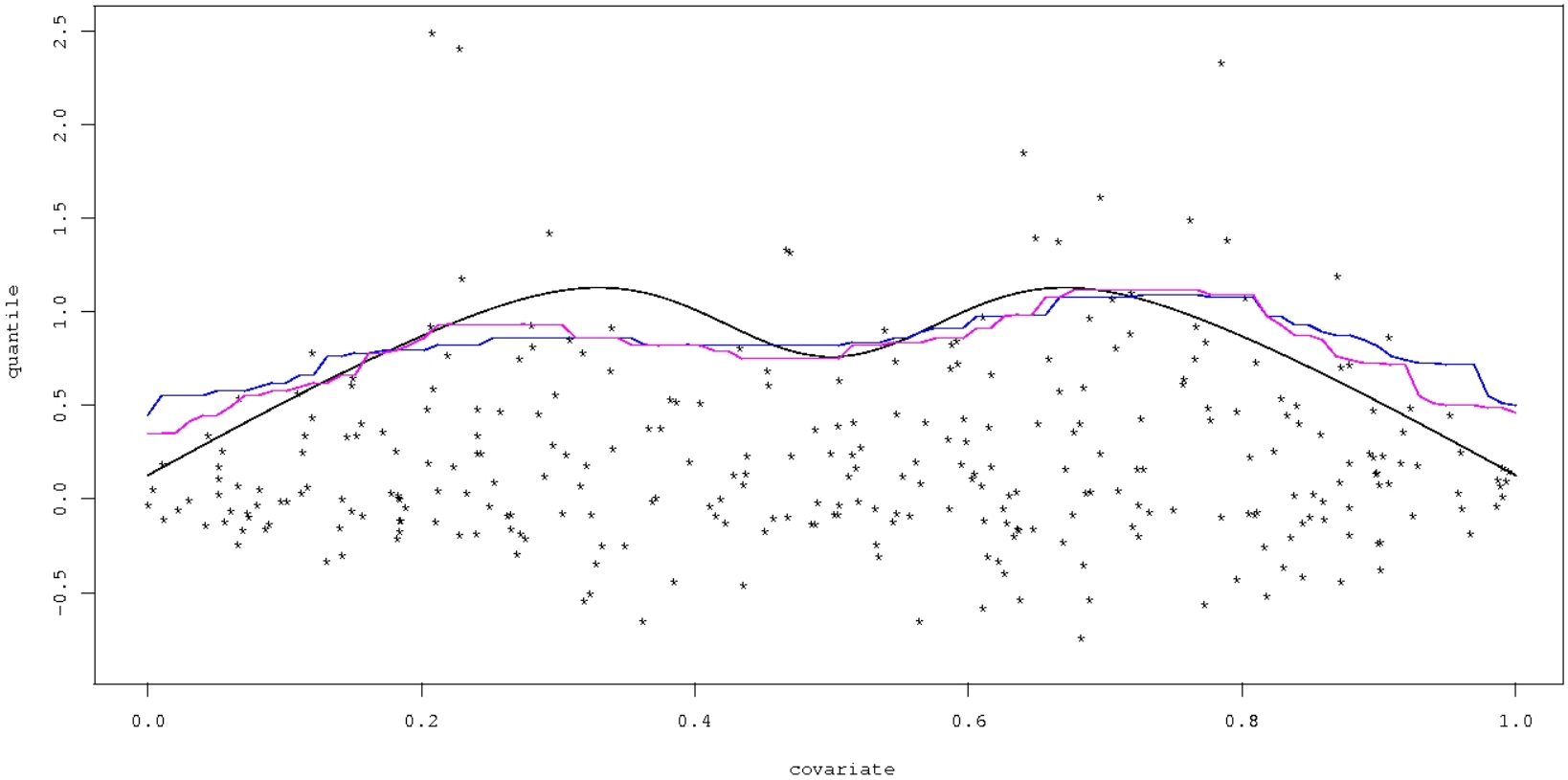
bleu : validation croisée, transparent : oracle.

9ème décile de l'erreur Δ



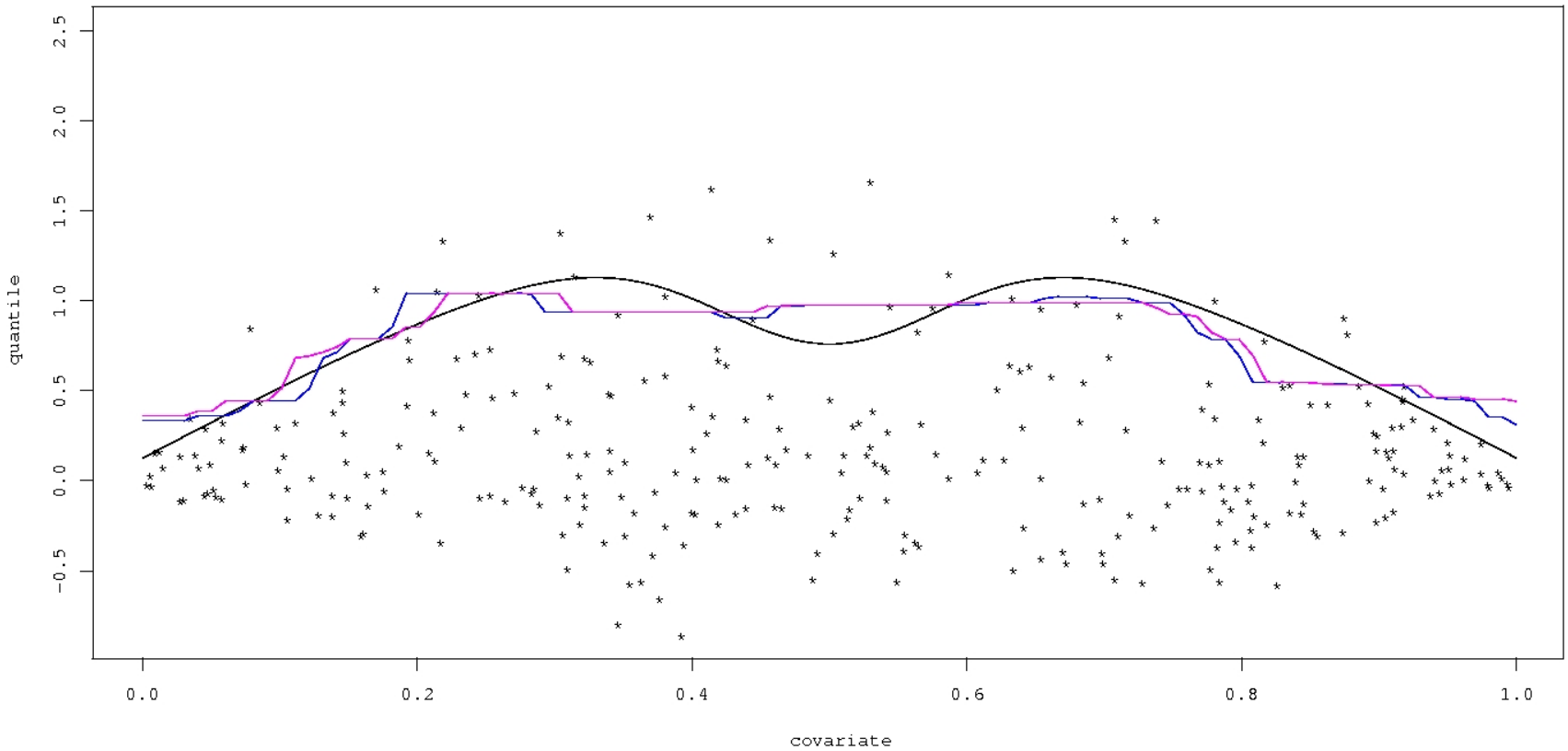
noir : vrai, bleu : validation croisée, rose : oracle.

médiane de l'erreur Δ



noir : vrai, bleu : validation croisée, rose : oracle.

1er décile de l'erreur Δ



noir : vrai, bleu : validation croisée, rose : oracle.

Perspectives

- Normalité asymptotique de l'estimateur à noyau de la fonction de survie dans un contexte plus général (sans hypothèse de domaine d'attraction).
- Extension des résutats de normalité asymptotiques multivariés à des résultats de convergence de processus (indexés par la covariable x ou l'ordre des quantiles α).
- Définition d'autres estimateurs de l'indice de queue conditionnel.

Bibliographie

Fonctions à variations régulières

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

Extrêmes non-conditionnels

- Hill, B.M. (1975) A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, **3**, 1163–1174.
- Pickands, J. (1975) Statistical inference using extreme order statistics. *The Annals of Statistics*, **3**, 119–131.
- Weissman, I. (1978). Estimation of parameters and large quantiles based on the k largest observations, *Journal of the American Statistical Association*, **73**, 812–815.

Bibliographie

Estimation fonctionnelle

- Samanta, T. (1989) Non-parametric estimation of conditional quantiles. *Statistics and Probability Letters*, **7**, 407–412.
- Berlinet, A., Gannoun, A., Matzner-Løber, E. (2001) Asymptotic normality of convergent estimates of conditional quantiles. *Statistics*, **35**, 139–169.
- Yao, Q. (1999) Conditional predictive regions for stochastic processes, *Technical report*, University of Kent at Canterbury.

Extrêmes conditionnels

- Gardes, L., Girard, S. (2010) Conditional extremes from heavy-tailed distributions: An application to the estimation of extreme rainfall return levels. *Extremes*, à paraître.