

Estimation non-paramétrique de quantiles extrêmes conditionnels

Stéphane Girard

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

Décembre 2010

en collaboration avec [Abdelaati Daouia](#), [Laurent Gardes](#) et d'après le travail de thèse d'[Alexandre Lekina](#)

- 1 Introduction
- 2 Estimation des petites probabilités conditionnelles
- 3 Estimation des quantiles extrêmes conditionnels
- 4 Illustration sur simulations

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 $\alpha \rightarrow 0$.

- Statistique fonctionnelle : une covariable $X \in E$ (espace métrique) 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.

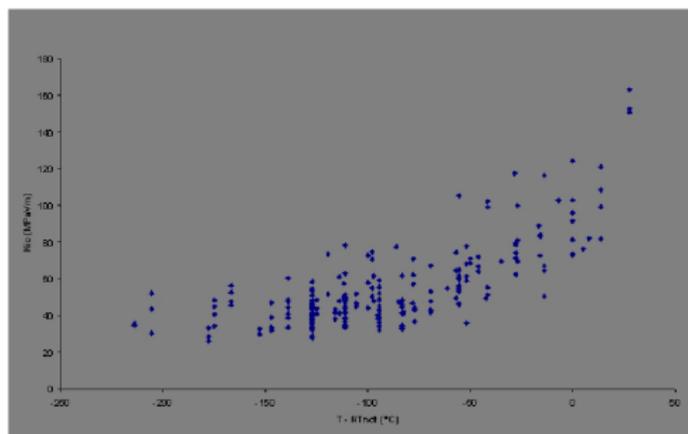
- Statistique des valeurs extrêmes “conditionnelles” : on s’intéresse aux **quantiles extrêmes conditionnels** définis par

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

quand $\alpha \rightarrow 0$.

Illustration : covariable unidimensionnelle ($E = \mathbb{R}$)

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

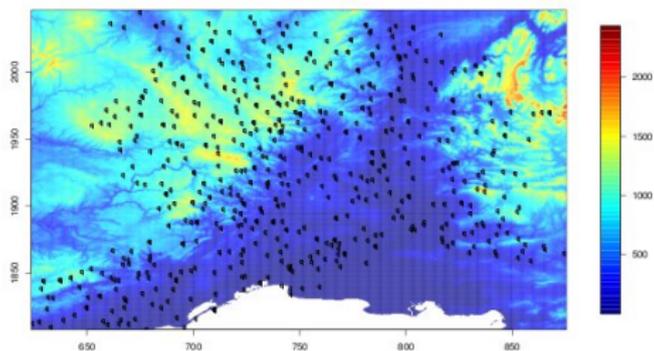


Y est la ténacité de la cuve (verticalement), X est la température (horizontalement).

Illustration : covariable tridimensionnelle ($E = \mathbb{R}^3$)

Données fournies par le Laboratoire des Transferts en Hydrologie et Environnement (LTHE) de Grenoble dans le cadre d'une ANR.

$X = \{\text{longitude, latitude, altitude}\}$, 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) , $i = 1, \dots, n$ de couples i.i.d.

- Estimer les **quantiles extrêmes conditionnels** définis par

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

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

- Estimer les **petites probabilités conditionnelles** définies par

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

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

- 1 Introduction
- 2 Estimation des petites probabilités conditionnelles
- 3 Estimation des quantiles extrêmes conditionnels
- 4 Illustration sur simulations

Principe

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

$$\hat{F}_n(y|x) = \frac{\sum_{i=1}^n K(d(x, X_i)/h) \mathbb{I}\{Y_i > y\}}{\sum_{i=1}^n K(d(x, X_i)/h)},$$

avec

- $\mathbb{I}\{.\}$ la fonction indicatrice,
- d une distance sur E ,
- $h = h_n$ une suite déterministe tq $h \rightarrow 0$ quand $n \rightarrow \infty$,
- K une fonction de support $[0, 1]$ telle que $0 < C_1 \leq K(t) \leq C_2$ pour tout $t \in [0, 1]$.

Hypothèse de domaine d'attraction

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

$$\bar{F}(y|x) = c(x)y^{-1/\gamma(x)} \exp\left(\int_1^y \frac{\varepsilon(u|x)}{u} du\right),$$

- $\gamma(\cdot)$ une fonction positive de la covariable x appelée **indice de queue conditionnel**,
- $c(\cdot)$ est une fonction positive de la covariable x ,
- $|\varepsilon(\cdot|x)|$ une fonction continue et décroissante vers 0.

Cela implique que $\bar{F}(\cdot|x)$ est à variations régulières d'indice $-1/\gamma(x)$ *i.e.* $\forall \lambda > 0$

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

Hypothèses de régularité

- **Conditions de Lipschitz.** Il existe des constantes positives κ_γ , κ_c et κ_ε telles que pour tout $(x, x') \in E \times E$,

$$\begin{aligned} \left| \frac{1}{\gamma(x)} - \frac{1}{\gamma(x')} \right| &\leq \kappa_\gamma d(x, x'), \\ |\log c(x) - \log c(x')| &\leq \kappa_c d(x, x'), \\ \sup_{u>1} |\varepsilon(u|x) - \varepsilon(u|x')| &\leq \kappa_\varepsilon d(x, x'). \end{aligned}$$

- **Notations.**

$B(x, h)$: boule de centre x et de rayon h ,

$\varphi_x(h) = \mathbb{P}(X \in B(x, h))$: probabilité de petite boule,

$\mu_x(h) = \mathbb{E}(K(x, X)/h)$.

Normalité asymptotique

Théorème

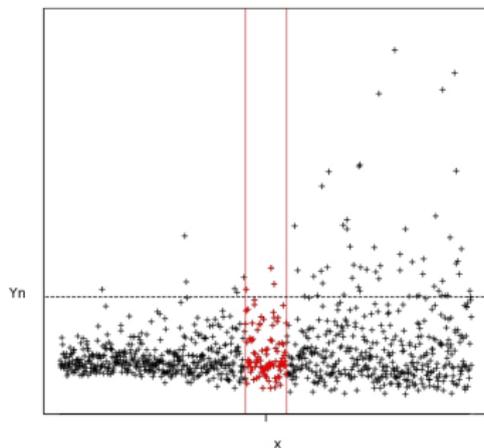
Soient $0 < a_1 < a_2 < \dots < a_J$, $J \in \mathbb{N}^*$ et $x \in E$ tel que $\varphi_x(h) > 0$. Si $y_n \rightarrow \infty$ tel que $n\varphi_x(h)\bar{F}(y_n|x) \rightarrow \infty$ et $n\varphi_x(h)(h \log y_n)^2 \bar{F}(y_n|x) \rightarrow 0$, alors

$$\left\{ \sqrt{n\mu_x(h)\bar{F}(y_n|x)} \left(\frac{\hat{F}_n(a_j y_n|x)}{\bar{F}(a_j y_n|x)} - 1 \right) \right\}_{j=1, \dots, J}$$

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

Condition $n\varphi_x(h)\bar{F}(y_n|x) \rightarrow \infty$

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



Si y_n est borné et $E = \mathbb{R}^p$, alors on retrouve la condition de normalité asymptotique classique : $nh^p \rightarrow \infty$.

Condition $n\varphi_x(h)(h \log y_n)^2 \bar{F}(y_n|x) \rightarrow 0$

- 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

$$\frac{1}{n\mu_x(h)\bar{F}(y_n|x)} \asymp \frac{1}{n\varphi_x(h)\bar{F}(y_n|x)}$$

- Si y_n est borné et $E = \mathbb{R}^p$, alors on retrouve la condition de normalité asymptotique classique : $nh^{p+2} \rightarrow 0$.

- 1 Introduction
- 2 Estimation des petites probabilités conditionnelles
- 3 Estimation des quantiles extrêmes conditionnels
- 4 Illustration sur simulations

Principe

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

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

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) quand $E = \mathbb{R}^p$ et par (Ferraty *et al.*, 2005) pour E quelconque.
- On pose $\sigma_n(x) = (n\mu_x(h)\alpha_n)^{-1/2}$.

Normalité asymptotique

Théorème

Soient $\tau_1 > \tau_2 > \dots > \tau_J > 0$, $J \in \mathbb{N}^*$ et $x \in E$ tel que $\varphi_x(h) > 0$.
Si $\alpha_n \rightarrow 0$ tel que $\sigma_n(x) \rightarrow 0$ et $\sigma_n^{-1}(x)h \log \alpha_n \rightarrow 0$, alors

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

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

Remarques sur la variance asymptotique

Si $E = \mathbb{R}^p$, elle est de l'ordre de

$$\frac{\gamma^2(x)}{n\alpha_n\mu_x(h)} \asymp \frac{\gamma^2(x)}{n\alpha_n\varphi_x(h)} \asymp \frac{\gamma^2(x)}{g(x)} \frac{1}{nh^p\alpha_n}.$$

- Rôle de l'indice de queue conditionnel. Un modèle équivalent : indice de queue 1 et probabilité de petite boule $\varphi_x(h)/\gamma^2(x)$.
- Facteur supplémentaire $1/\alpha_n$ par rapport au cas $\alpha_n = \alpha$ fixé (Berlinet *et al.*, 2001, Théorème 6.4).
- On peut choisir $h = \eta_n(n\alpha_n \log^2(\alpha_n))^{-1/(p+2)}$ où $\eta_n \rightarrow 0$. 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

Si $E = \mathbb{R}^p$, les deux conditions $nh^p\alpha_n \rightarrow \infty$ et $nh^{p+2} \log^2(\alpha_n)\alpha_n \rightarrow 0$ entraînent

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

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.

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

Normalité asymptotique

Corollaire

Soit $x \in E$ tel que $\varphi_x(h) > 0$. Si $\alpha_n \rightarrow 0$ tel que $\sigma_n(x) \rightarrow 0$, $\sigma_n^{-1}(x)h \log(\alpha_n) \rightarrow 0$ et $\sigma_n^{-1}(x)\varepsilon(q(2\alpha_n|x)|x) \rightarrow 0$, alors

$$\sigma_n^{-1}(x)(\hat{\gamma}_n^p(x) - \gamma(x))$$

est asymptotiquement gaussien centré de variance

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

Condition $\sigma_n^{-1}(x)\varepsilon(q(2\alpha_n|x)|x) \rightarrow 0$

Rappel : modèle

$$\bar{F}(y|x) = c(x)y^{-1/\gamma(x)} \exp\left(\int_1^y \frac{\varepsilon(u|x)}{u} du\right).$$

Lorsque $y \rightarrow \infty$,

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

et $|\varepsilon(\cdot|x)|$ étant décroissante à l'infini :

$$\left| \lambda^{1/\gamma(x)} \frac{\bar{F}(\lambda y|x)}{\bar{F}(y|x)} - 1 \right| \leq (1 + o(1)) |\varepsilon(y|x)| \log(\lambda).$$

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

Estimation de l'indice de queue conditionnel

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

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

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

Normalité asymptotique

Corollaire

Soit $x \in E$ tel que $\varphi_x(h) > 0$. Si $\alpha_n \rightarrow 0$ tel que $\sigma_n(x) \rightarrow 0$, $\sigma_n^{-1}(x)h \log(\alpha_n) \rightarrow 0$ et $\sigma_n^{-1}(x)\varepsilon(q(\alpha_n|x)|x) \rightarrow 0$, alors

$$\sigma_n^{-1}(x)(\hat{\gamma}_n^H(x) - \gamma(x))$$

est asymptotiquement gaussien centré de variance $\gamma^2(x)V_J$ avec

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

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.

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

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

Soit $x \in E$. Si

- $\alpha_n \rightarrow 0$ tel que $\sigma_n(x) \rightarrow 0$ et $\sigma_n^{-1}(x)h \log(\alpha_n) \rightarrow 0$,
- $\beta_n/\alpha_n \rightarrow 0$
- $\sigma_n^{-1}(x)(\hat{\gamma}_n(x) - \gamma(x)) \xrightarrow{d} \mathcal{N}(0, v^2(x))$,

alors

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

- 1 Introduction
- 2 Estimation des petites probabilités conditionnelles
- 3 Estimation des quantiles extrêmes conditionnels
- 4 **Illustration sur simulations**

Illustration sur simulations

- Echantillon $\{(Y_i, X_i), i = 1, \dots, n\}$ 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.*

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

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

- **But** : estimation du quantile extrême conditionnel $q(\alpha_n|x) = (-\log(1 - \alpha_n))^{-\gamma(x)}$ d'ordre $\alpha_n = 5 \log(n)/n$.

Choix de la fenêtre de lissage

- Approche préconisée : **validation croisée** (Yao, 1999)

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

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

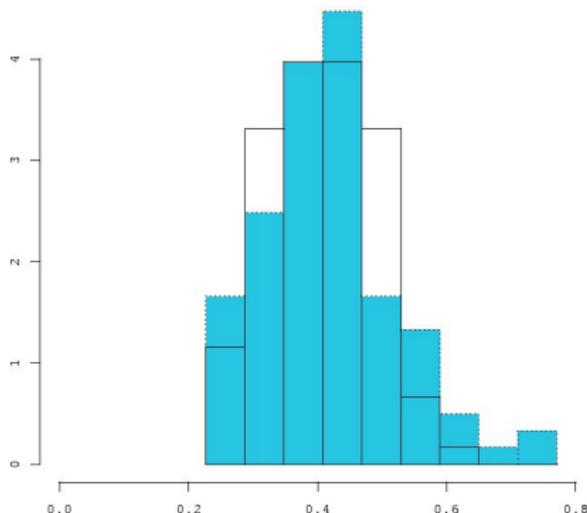
- Approche de référence : **oracle**. Minimisation de la distance Δ entre quantile estimé et quantile théorique :

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

avec

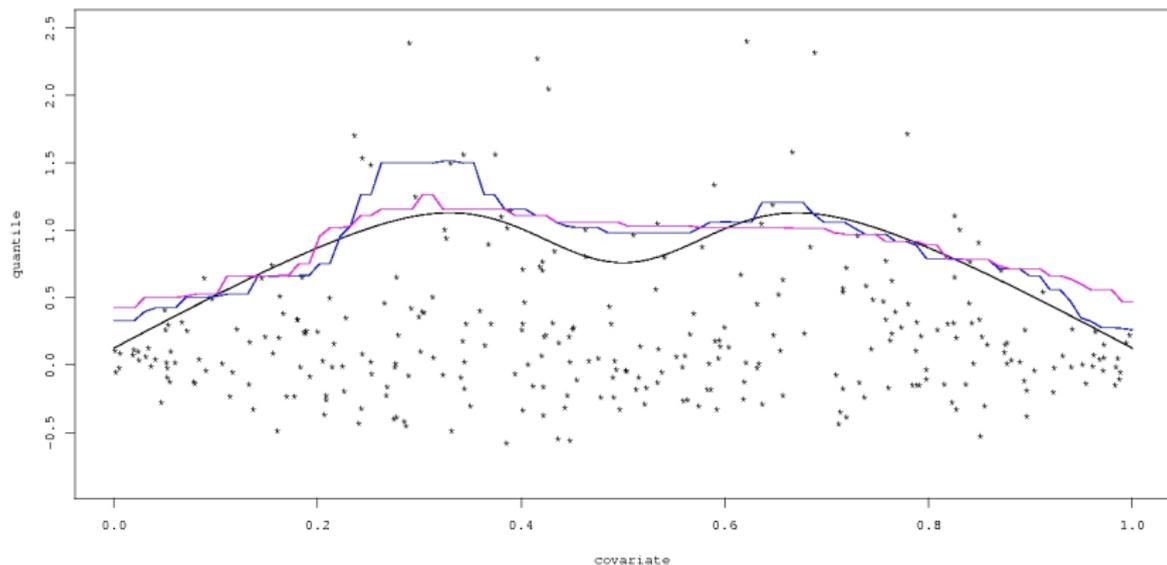
$$\Delta(q(\alpha_n|\cdot), \hat{q}_n(\alpha_n|\cdot)) = \left\{ \frac{1}{L} \sum_{\ell=1}^L (\hat{q}_n(\alpha_n|t_\ell) - q(\alpha_n|t_\ell))^2 \right\}^{1/2}.$$

Histogramme des erreurs (calculé sur 100 réplifications)

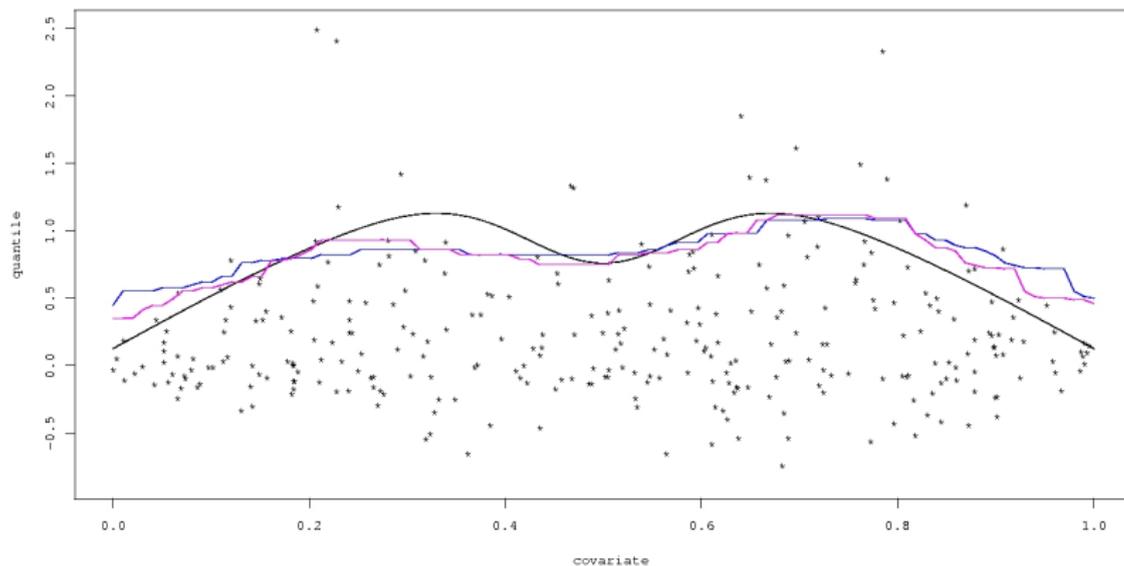


bleu : validation croisée, transparent : oracle.

9ème décile de l'erreur Δ

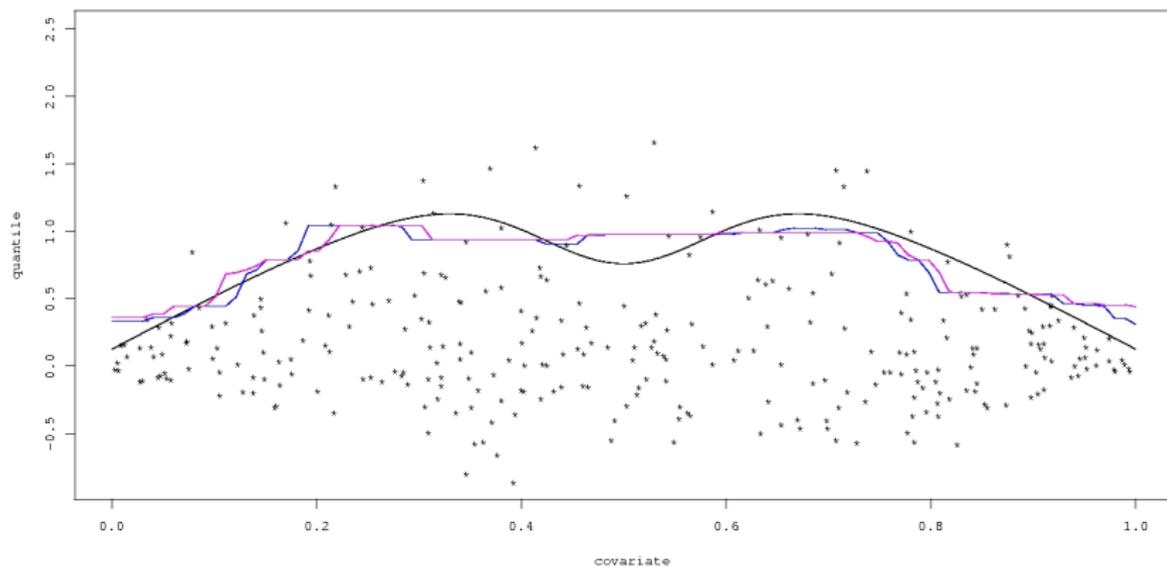


noir : vrai, bleu : validation croisée, rose : oracle.

médiane de l'erreur Δ 

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

1er décile de l'erreur Δ

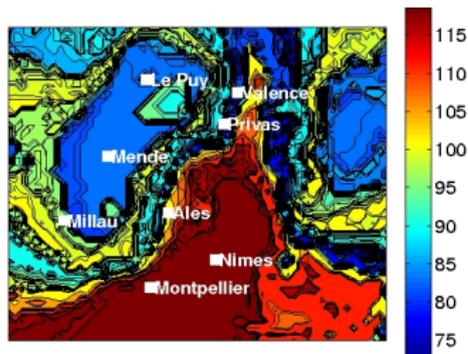


noir : vrai, bleu : validation croisée, rose : oracle.

Retour sur l'illustration ($E = \mathbb{R}^3$)

Données fournies par le Laboratoire des Transferts en Hydrologie et Environnement (LTHE) de Grenoble dans le cadre d'une ANR.

$X = \{\text{longitude, latitude, altitude}\}$, Y : hauteur de pluie.



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)

Perspectives

- Normalité asymptotique quel que soit le domaine d'attraction.
- Extension des résultats 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

- Berlinet, A., Gannoun, A., Matzner-Løber, E. (2001) Asymptotic normality of convergent estimates of conditional quantiles. *Statistics*, **35**, 139–169.
- Ferraty, F., Rabhi, A., Vieu, P. (2005) Conditional quantiles for dependent functional data with application to the climatic El Nino phenomenon, *Sankhya: The Indian Journal of Statistics*, **67**(2), 378–398.
- Samanta, T. (1989) Non-parametric estimation of conditional quantiles. *Statistics and Probability Letters*, **7**, 407–412.
- Yao, Q. (1999) Conditional predictive regions for stochastic processes, *Technical report*, University of Kent at Canterbury.

Bibliographie

Extrêmes conditionnels

- Gardes, L., Girard, S. (2010) Conditional extremes from heavy-tailed distributions: An application to the estimation of extreme rainfall return levels. *Extremes*, **13**(2), 177–204.
- Daouia, A., Gardes, L., Girard, S., Lekina, A. (2010) Kernel estimators of extreme level curves, *Test*, to appear.