Functional kernel estimators of conditional extreme quantiles

Stéphane Girard

Inria Grenoble Rhône-Alpes, équipe-projet Mistis
http://mistis.inrialpes.fr/people/girard/

Juin 2012

en collaboration avec Laurent Gardes, Université de Strasbourg.

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 $X \in E$ (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.

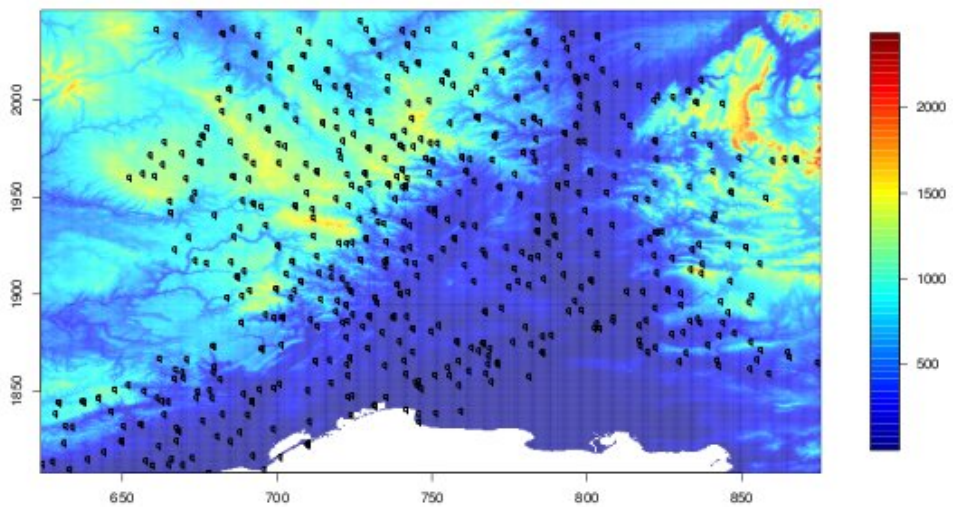
- On s'intéresse aux quantiles extrêmes conditionnels définis par

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

quand $\alpha \to 0$.

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.
$X = \{\text{longitude, lattitude, altitude}\}$, 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 $\alpha_n \to 0$ lorsque $n \to \infty$.

- Estimer les petites probabilités conditionnelles définies par

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

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

Principe

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

$$\hat{\bar{F}}_n(y|x) = \sum_{i=1}^{n} K(d(x,X_i)/h) Q((Y_i - y)/\lambda) \Big/ \sum_{i=1}^{n} K(d(x,X_i)/h),$$

avec

- d une semi-métrique sur E,
- $h = h_n$ et $\lambda = \lambda_n$ deux suites tq $h \to 0$ quand $n \to \infty$,
- K une fonction de support $[0,1]$ telle que $0 < C_1 \leq K(t) \leq C_2$ pour tout $t \in [0,1]$,
- Q est une fonction de répartition associée à une densité à support compact.

Hypothèse de domaine d'attraction

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

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

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

Cela implique que $\bar{F}(.|x)$ est à variations régulières d'indice $-1/\gamma(x)$ *i.e.* $\forall t > 0$

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

Hypothèses de régularité

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

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

- 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,
  $\mu_x^{(\tau)}(h) = \mathbb{E}(K^\tau(x, X)/h))$, pour tout $\tau > 0$.

Normalité asymptotique

Théorème 1

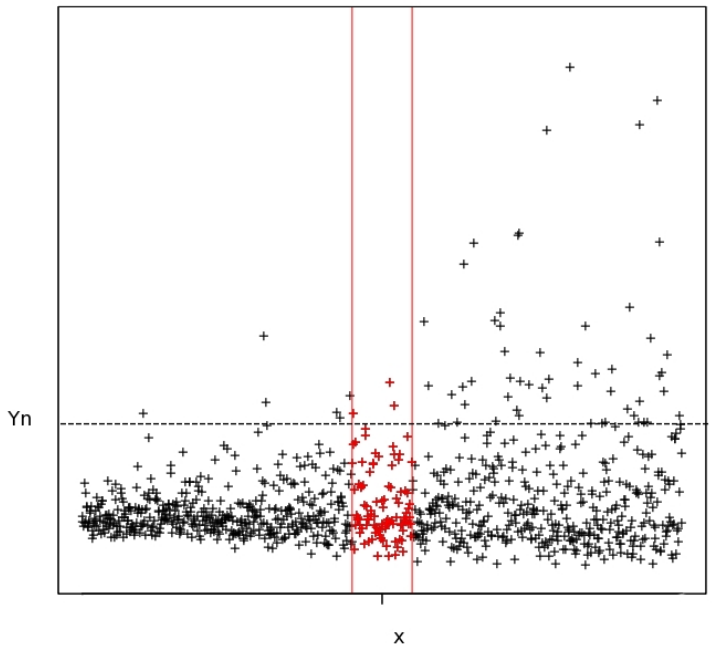
Soient $0 < a_1 < a_2 < \cdots < a_J$, $J \in \mathbb{N}^*$ et $x \in E$ tq $\varphi_x(h) > 0$.
Si $y_n \to \infty$, $n\varphi_x(h)\bar{F}(y_n|x) \to \infty$ et
$n\varphi_x(h)\bar{F}(y_n|x)(h \log y_n \vee \lambda/y_n)^2 \to 0$, alors

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

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

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

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



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

Condition $n\varphi_x(h)\bar{F}(y_n|x)(\lambda/y_n \vee h\log y_n)^2 \to 0$

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

$$(\lambda/y_n \vee h\log y_n)^2,$$

soit négligeable devant la variance, de l'ordre de

$$\Lambda_n^2(x) = \frac{1}{n\bar{F}(y_n|x)} \frac{\mu_x^{(2)}(h)}{(\mu_x^{(1)}(h))^2} \asymp \frac{1}{n\bar{F}(y_n|x)\varphi_x(h)}$$

- Si y_n est borné et $E = \mathbb{R}^p$, alors on retrouve la condition de normalité asymptotique classique : $nh^p(h \vee \lambda)^2 \to 0$.

Principe

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

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

  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

$$\sigma_n^2(x) = \frac{1}{n\alpha_n} \frac{\mu_x^{(2)}(h)}{(\mu_x^{(1)}(h))^2}.$$

Normalité asymptotique

Théorème 2

Soient $\tau_1 > \tau_2 > \cdots > \tau_J > 0$, $J \in \mathbb{N}^*$ et $x \in E$ tel que $\varphi_x(h) > 0$. Si $\alpha_n \to 0$ tel que $\sigma_n(x) \to 0$ et $\sigma_n^{-1}(x)(h \log \alpha_n \vee \lambda / q(\alpha_n|x)) \to 0$, alors

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

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

Remarques sur la variance asymptotique

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

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

- 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 $\eta_n \to 0$. 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 non conditionnels.

Remarque sur l'ordre des quantiles extrêmes

Si $E = \mathbb{R}^p$, 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".

Estimateur de Weissman fonctionnel

Adaptation de l'estimateur de Weissman (Weissman, 1978) au cas fonctionnel.

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

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 très extrêmes.

Normalité asymptotique

Théorème 3

Soit $x \in E$. Si

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

alors

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

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

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 $y \longrightarrow \infty$,

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

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

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

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

Estimation de l'indice de queue fonctionnel

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

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

Corollaire

Sous les hypothèses du Théorème 2, et si de plus $\sigma_n^{-1}(x)\varepsilon(q(\alpha_n|x)|x) \to 0$, alors $\sigma_n^{-1}(x)(\hat{\gamma}_n^{\mathrm{P}}(x) - \gamma(x))$ est asymptotiquement gaussien centré de variance

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

.

Estimation de l'indice de queue fonctionnel

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

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

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

Corollaire

Sous les hypothèses du Théorème 2, et si de plus $\sigma_n^{-1}(x)\varepsilon(q(\alpha_n|x)|x) \to 0$, alors $\sigma_n^{-1}(x)(\hat{\gamma}_n^{\mathrm{H}}(x) - \gamma(x))$ est asymptotiquement gaussien centré de variance $\gamma^2(x)V_J$ avec

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

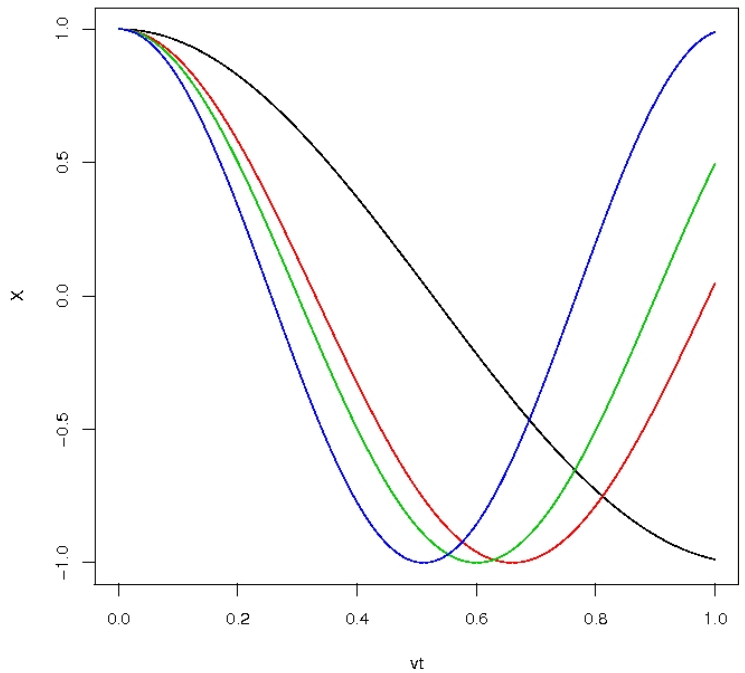
Illustration sur simulations

$N = 50$ réplications d'un échantillon de taille $n = 500$ d'un couple (X, Y).

- $X \in E = L^2[0,1]$ est défini par $X(t) = \cos(2\pi Z t)$ pour tout $t \in [0,1]$ où Z est uniforme sur $[1/4, 1]$.
- Y sachant X suit une loi de Burr $\bar{F}(y|X) = (y^2+1)^{-1/2\gamma(X)}$ avec $\gamma(X) = 2\|X\|_2^2 - 3/4$ et

$$\|X\|_2^2 = \int_0^1 X^2(t)dt = \frac{1}{2}\left(1 + \frac{\sin(4\pi Z)}{4\pi Z}\right).$$

Quatre réalisations de fonctions $X(.)$



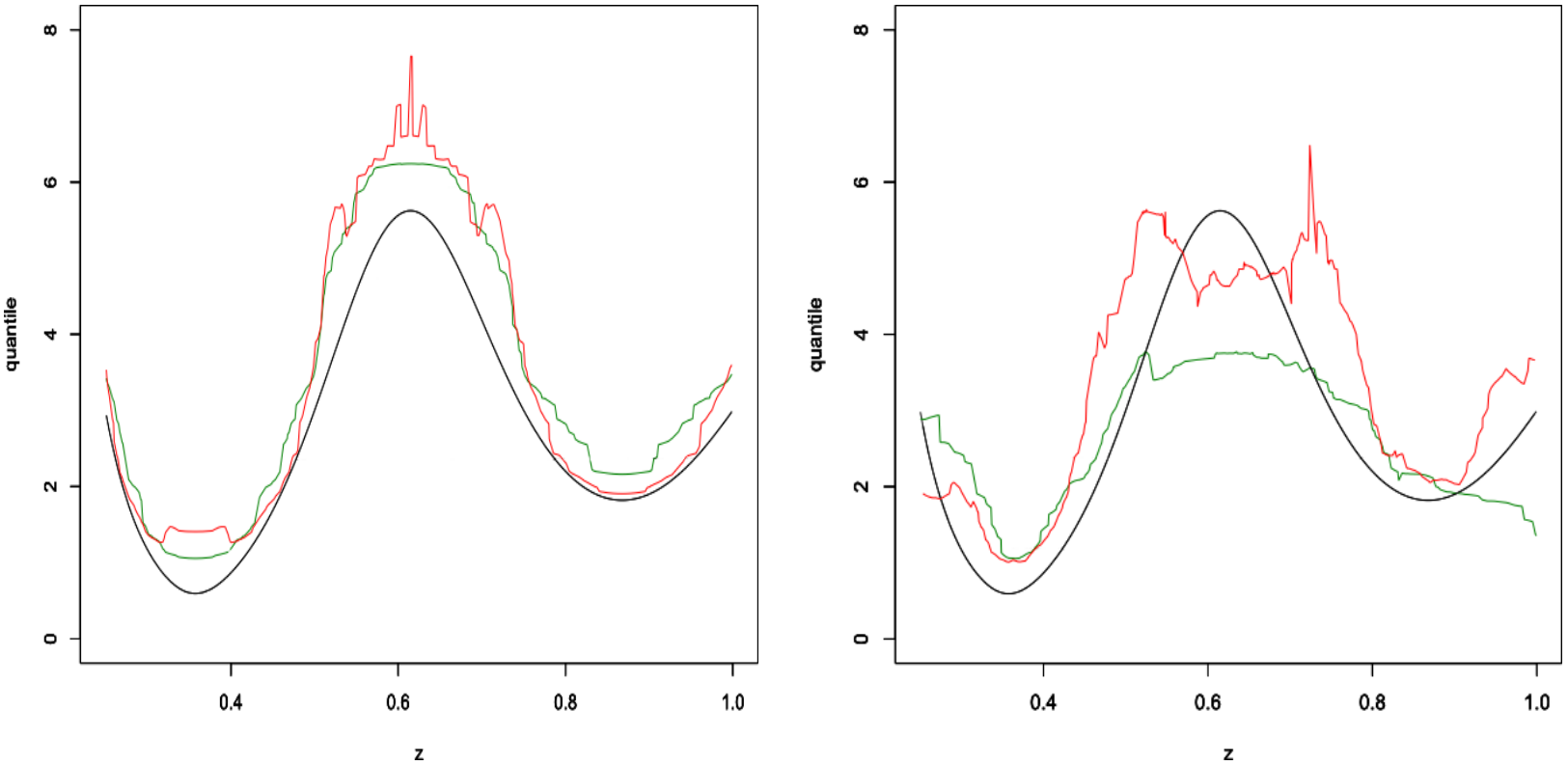
- **But :** Estimation du quantile $q(\beta_n|x)$ avec $\beta_n = 5/n$.
- **Outils :** Estimateur de Weissman fonctionnel combiné à l'estimateur de Hill fonctionnel.
  - Poids : $\tau_j = (1/j)^s$ avec $s \in \{1, 2, 3, 10\}$.
  - Ordres du quantile peu extrême : $\alpha_n = c \log(n)/n$ avec $c \in \{5, 10, 15, 20\}$.
  - Deux choix de semi-métriques : $d_X(s,t) = \|s - t\|_2$ et $d_Z(s,t) = \left| \|s\|_2^2 - \|t\|_2^2 \right|$.
  - Noyaux : $K(t) = (1.9 - 1.8t)\mathbb{I}\{t \in [0,1]\}$ et noyau triangulaire.
  - Paramètre de lissage : validation croisée (Yao, 1999)

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

Résultats médians

		$c=5$	$c=10$	$c=15$	$c=20$
$s=1$	$d=d_X$	13457	854	756	913
	$d=d_Z$	747	527	589	618
$s=2$	$d=d_X$	1792	531	435	472
	$d=d_Z$	420	**360**	**347**	309
$s=3$	$d=d_X$	1225	**450**	**396**	352
	$d=d_Z$	329	304	261	242
$s=10$	$d=d_X$	∞	616	359	242
	$d=d_Z$	67267	116	464	840

Résultats médians



Gauche/droite : influence de la métrique d_Z/d_X,
Rouge/vert : influence de α_n et τ_js.

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 (dimension finie)

- 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. (2011) Kernel estimators of extreme level curves, *Test*, **20**(2), 311–333.