ESTIMATION DE COURBES DE NIVEAUX EXTRÊMES POUR DES DISTRIBUTIONS À QUEUES LOURDES

Alexandre LEKINA

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

En collaboration avec A. DAOUIA[1], L. GARDES & S. GIRARD

Mai 2010

LABORATOIRE JEAN KUNTZMANN
MATHÉMATIQUES APPLIQUÉES - INFORMATIQUE

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

[1] GREMAQ, Université de Toulouse.

But

- Soient $\{(X_i, Y_i), i = 1, \ldots, n\}$ des copies indépendantes du couple aléatoire $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$.
- Y est une variable d'intérêt associée à une covariable X.
- Estimer pour tout $x \in \mathbb{R}^d$ et pour tout $\alpha_n \to 0$, les courbes de niveaux extrêmes définies comme les graphes de fonctions $x \in \mathbb{R}^d \mapsto q(\alpha_n|x) \in \mathbb{R}$ vérifiant

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

lorsque la fonction de répartition conditionnelle de Y sachant $X = x$ est à queue lourde d'indice $-1/\gamma(x)$, i.e

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

avec

- $\gamma(.)$ une fonction inconnue et positive de la covariable appelée l'"indice de queue conditionnel".
- $\ell(.|x)$ une "fonction à variations lentes", i.e pour tout $\lambda > 0$,

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

Estimateur: l'inverse généralisé de l'estimateur de la fonction de survie conditionnelle

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

Nécessite d'estimer la probabilité $\bar{F}(y_n|x)$ lorsque $y_n \to \infty$ (quand $n \to \infty$).

Estimateur à noyau de $\bar{F}$, (Collomb (1976))

$$\hat{\bar{F}}_n(y|x) = \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right) \mathbf{1}_{\{Y_i > y\}} \Bigg/ \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right)$$

- $\mathbf{1}_{\{.\}}$ est la fonction indicatrice.
- La fonction $K(.)$ appelée noyau est positive, bornée, intégrable et à support compact $S \subseteq \mathbb{R}^d$.
- h_n est une suite non aléatoire telle que $h_n \to 0$ quand $n \to \infty$.

Normalité asymptotique $\hat{\bar{F}}_n(y_n|x)$, (Daouia, Gardes, Girard & Lekina (2010))

Soit g la densité de X. Sous des conditions de régularité de type Lipschitz sur $\gamma(x)$, $g(x)$ et $\log \ell(.|x)/\log(.)$. Si

- $y_n \to \infty$ tq $nh_n^d \bar{F}_n(y_n|x) \to \infty$ et $nh_n^{d+2} \bar{F}_n(y_n|x) \log^2(y_n) \to 0$ qd $n \to \infty$,
- $\{a_j, j = 1, \ldots, J\}$ est une suite strictement positive et croissante,

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

$$\left\{ \sqrt{nh_n^d \bar{F}(y_n|x)} \left(\frac{\hat{\bar{F}}_n(a_j y_n|x)}{\bar{F}(a_j y_n|x)} - 1 \right) \right\}_{\{j=1,\ldots,J\}} \xrightarrow{\mathcal{L}} \mathcal{N}\left(0_{\mathbb{R}^J}, \frac{\|K\|_2^2}{g(x)} C(x) \right),$$

où $C_{j,j'}(x) = a_{j \wedge j'}^{1/\gamma(x)}$ pour $(j, j') \in \{1, \ldots, J\}^2$.

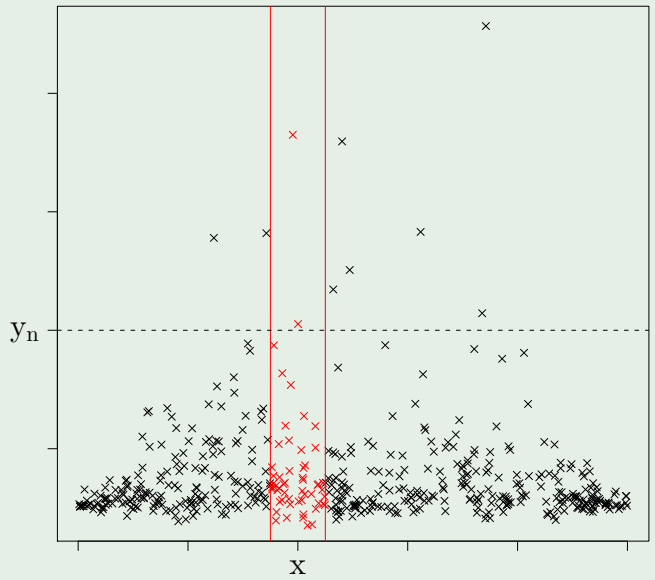
Interprétation de la condition $nh_n^{d+2} \bar{F}_n(y_n|x) \log^2(y_n) \to 0$

- Condition pour que le carré du biais asymptotique, de l'ordre de $(h_n \log y_n)^2$ soit négligeable devant la variance asymptotique, de l'ordre de $1/\{nh_n^d \bar{F}_n(y_n|x)\}$.
- Si y_n est borné, alors on retrouve la condition de normalité asymptotique classique: $nh_n^{d+2} \to 0$.

Interprétation de la condition $nh_n^d \bar{F}_n(y_n|x) \to \infty$

- Condition Nécessaire et Suffisante pour qu'il y ait presque sûrement au moins un point dans la région $B(x, h_n) \times [y_n, +\infty[$ de $\mathbb{R}^{d+1}$.
- Si y_n est borné, alors on retrouve la condition de normalité asymptotique classique: $nh_n^d \to \infty$.

Illustration géométrique de la condition $nh_n^d \bar{F}_n(y_n|x) \to \infty$



Normalité asymptotique de $\hat{q}_n(\alpha_n|x)$, (Daouia, Gardes, Girard & Lekina (2010))

Soit g la densité de X. Sous des conditions de régularité de type Lipschitz sur $\gamma(x)$, $g(x)$ et $\log \ell(.|x)/\log(.)$. Si

- $\ell(.|x)$ est normalisée (voir Bingham, Goldie et Teugels (1987)),
- $\alpha_n \to 0$ telle que $nh_n^d\alpha_n \to \infty$ et $nh_n^{d+2}\alpha_n \log^2(\alpha_n) \to 0$ quand $n \to \infty$,
- $\{\tau_j, j = 1, \ldots, J\}$ est une suite strictement positive et décroissante,

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

$$\left\{ \sqrt{nh_n^d\alpha_n} \left(\frac{\hat{q}_n(\tau_j\alpha_n|x)}{q(\tau_j\alpha_n|x)} - 1 \right) \right\}_{\{j=1,\ldots,J\}} \xrightarrow{\mathcal{L}} \mathcal{N}\left(0_{\mathbb{R}^J}, \gamma^2(x)\frac{\|K\|_2^2}{g(x)}\Sigma \right),$$

où $\Sigma_{j,j'}(x) = 1/\tau_{j\wedge j'}$ pour $(j, j') \in \{1, \ldots, J\}^2$.

Remarques :

- La variance asymptotique est inversement proportionnelle à $nh_n^d\alpha_n$ et proportionnelle à $\gamma^2(x)$.
- $nh_n^d\alpha_n \to \infty$ et $nh_n^{d+2}\alpha_n \log^2(\alpha_n) \to 0$ entrainent $\frac{n\alpha_n}{\log^d(1/\alpha_n)} \to \infty$ qui implique $\alpha_n > \frac{\log^d(n)}{n}$: on ne peut pas estimer des quantiles très extrêmes.

Application 1 : Estimateur à noyau de type Hill (voir Hill (1975))

$$\hat{\gamma}_n^{\mathrm{H}}(x) = \sum_{j=1}^{J} \log\left(\hat{q}_n(\tau_j \alpha_n | x)/\hat{q}_n(\alpha_n|x)\right) \Big/ \sum_{j=1}^{J} \log\left(1/\tau_j\right) \quad \text{avec } J > 1.$$

Condition du second ordre

La fonction $|\varepsilon(y|x)| \overset{def}{=} \left| y \dfrac{\ell'(y|x)}{\ell(y|x)} \right|$ est asymptotiquement décroissante et tend vers 0 lorsque $y \to \infty$.

Normalité asymptotique de $\hat{\gamma}_n^{\mathrm{H}}(x)$, (Daouia, Gardes, Girard & Lekina (2010))

Si $\sqrt{nh_n^d \alpha_n} \to \infty$ et $\sqrt{nh_n^d \alpha_n}\varepsilon(q(\alpha_n|x)|x) \to 0$ quand $n \to \infty$ alors,

$$\sqrt{nh_n^d \alpha_n}\left(\hat{\gamma}_n^{\mathrm{H}}(x) - \gamma(x)\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{V_J \|K\|_2^2}{g(x)}\gamma^2(x)\right),$$

où $V_J = \left(\sum_{j=1}^{J} \frac{2(J-j)+1}{\tau_j} - J^2\right) \Big/ \left(\sum_{j=1}^{J} \log\left(\tau_1/\tau_j\right)\right)^2$.

Quelques exemples de suites de poids

- Si $\tau_j = 1/j$, alors V_J est minimum pour $J = 9$ et $V_9 \simeq 1.245$.
- Si $\tau_j = (1/j)^{j/J}$, alors V_J est minimum pour $J = 15$ et $V_{15} \simeq 1.117$.

Application 2 : Un estimateur à noyau pour estimer les quantiles très extrêmes

Adapter l'estimateur de 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 $\hat{q}_n(\alpha_n|x)$ est l'estimateur précédent et $\hat{\gamma}_n(x)$ un estimateur de $\gamma(x)$.

Loi asymptotique de $\hat{q}_n^{\mathrm{W}}(\beta_n|x)$, (Daouia, Gardes, Girard & Lekina (2010))

- $\alpha_n \to 0$ telle que $nh_n^d\alpha_n \to \infty$ et $nh_n^{d+2}\alpha_n \log^2(\alpha_n) \to 0$ quand $n \to \infty$.
- $\sqrt{nh_n^d\alpha_n}\,(\hat{\gamma}_n(x) - \gamma(x)) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, v^2(x)\right)$ avec $v(x) > 0$.
- Si $\beta_n/\alpha_n \to 0$ alors, pour tout $x \in \mathbb{R}^d$ tel que $g(x) > 0$,

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

Experience numérique

- On génère $m = 100$ réplications d'un échantillon $\{(X_i, Y_i), i = 1, \ldots, n\}$ de taille $n = 500$ suivant la loi du couple $(X, Y) \in \mathbb{R} \times \mathbb{R}$ où $X \sim U[0, 1]$ et dont le quantile conditionnel de Y sachant $X = x$ est donné par

$$q(\alpha_n|x) = \left(\alpha_n^{\rho(x)} - 1\right)^{-\gamma(x)/\rho(x)} \text{ où on fixe } \rho(x) = -2 \quad \text{(Loi de Burr)}.$$

- But : estimer le quantile extrême conditionnel d'ordre $1/2n$.
- On estime $q(1/2n|.)$ par $\hat{q}_n^{\mathrm{W}}(1/2n|.)$.
- On utilise un noyau d'expression

$$K(x) = \frac{15}{16}\left(1 - x^2\right)^2 \mathbf{1}_{\{|x| \leq 1\}} \quad \text{(noyau biquadratique)}.$$

- On choisit le paramètre de lissage par validation croisée suivant le critère

$$h_{cv} = \arg \min_{h \in [0.02, 1/4]} \sum_{i=1}^{n} \sum_{j=1}^{n} \left\{ \mathbf{1}_{\{Y_i \geq Y_j\}} - \hat{\bar{F}}_{n,-i}\left(Y_j|X_i\right) \right\}^2 \quad \text{(Yao (1999))}$$

où $\hat{\bar{F}}_{n,-i}$ est l'estimateur de $\bar{F}$ calculé à partir de l'échantillon $\{(X_k, Y_k), k = 1, \ldots, n\}$ privé de sa ième observation (X_i, Y_i).

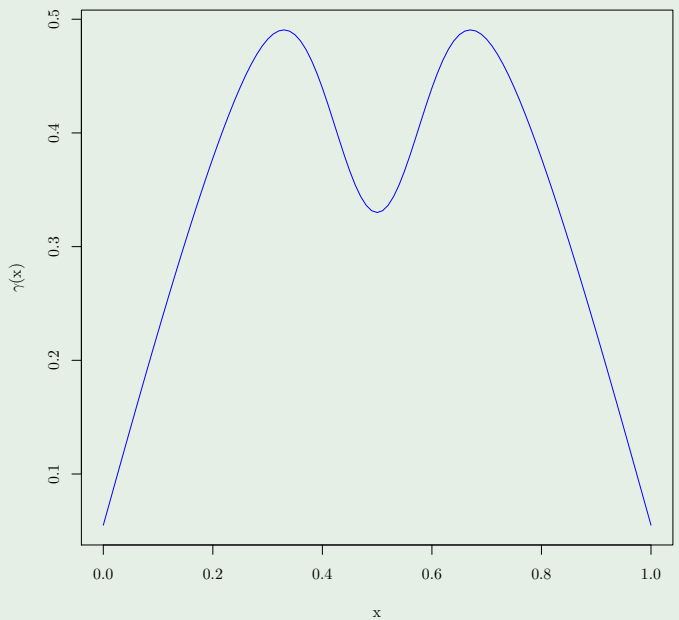
Experience numérique

- La fonction indice de queue conditionnelle est définie par

$$x \in [0,1] \mapsto \gamma(x) = \frac{1}{2}\left(\frac{1}{10} + \sin(\pi x)\right)\left(\frac{11}{10} - \frac{1}{2}\exp\left(-64\,(x-1/2)^2\right)\right).$$

- Pour estimer $\gamma(x)$ on pose $\tau_j = (1/j)^{j/J}$ et on fixe $J = 15$.

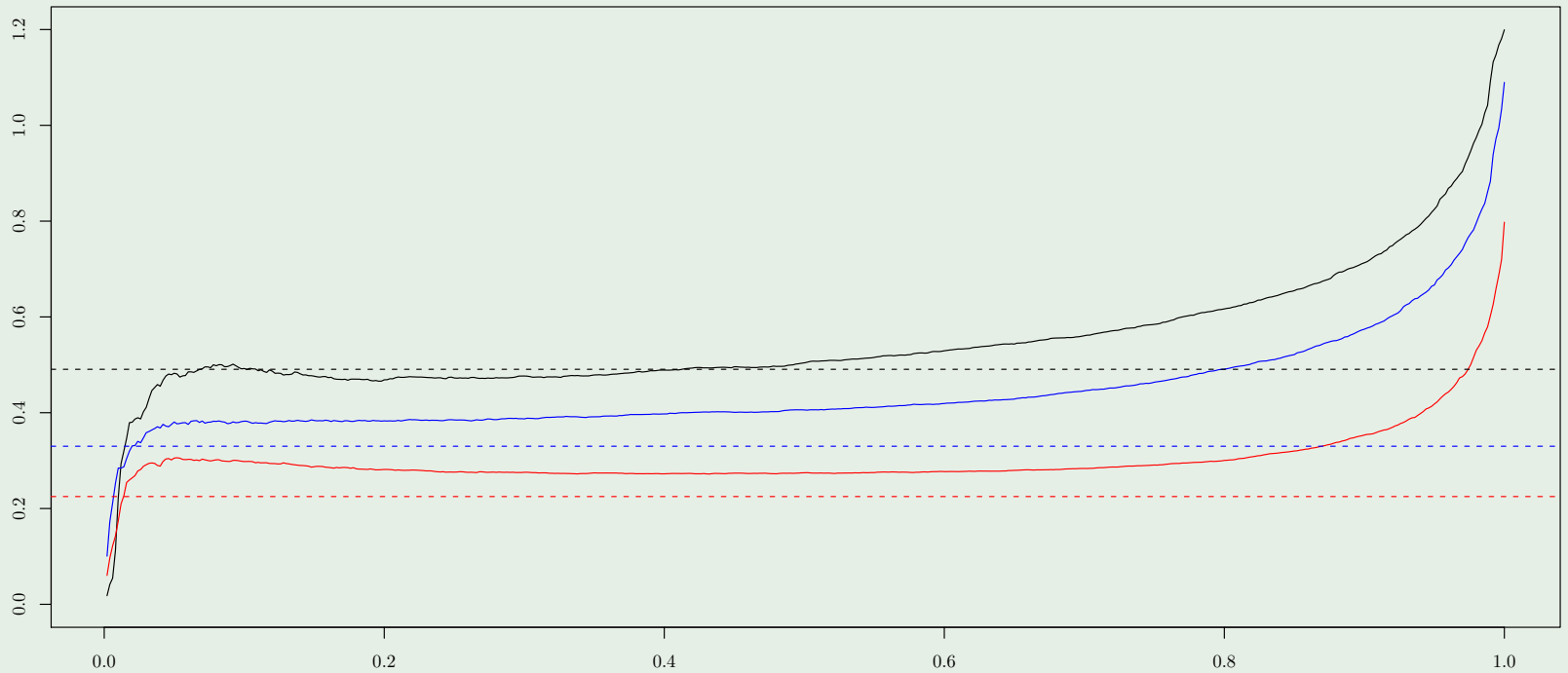
La fonction $\gamma(x)$



Graphique des moyennes de Hill en fonction de α_n.

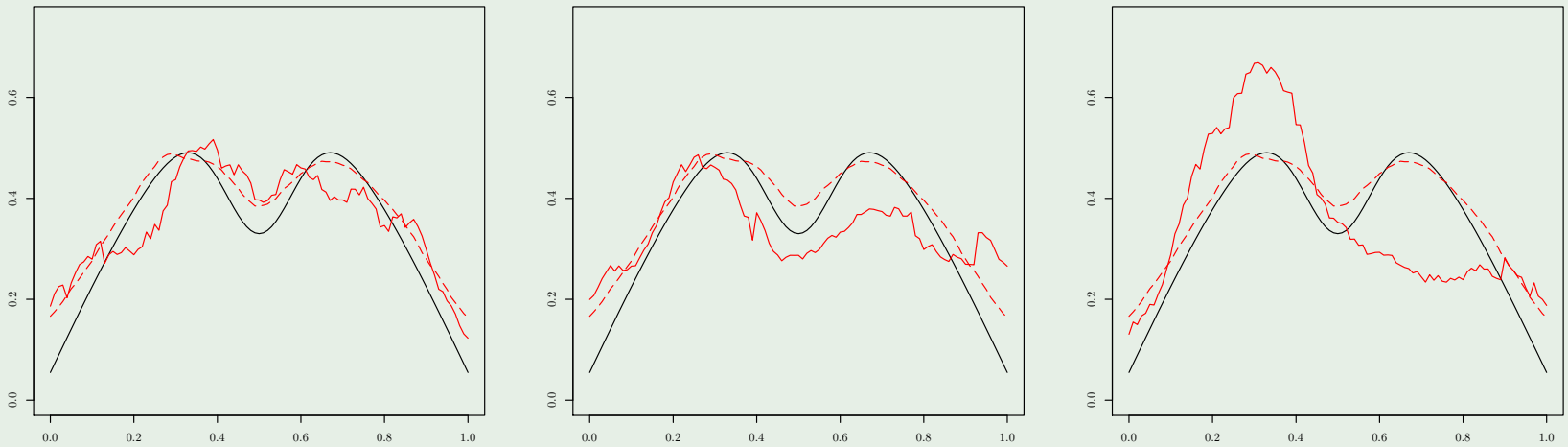
— $x = 0.68$ — $x = 0.51$ — $x = 0.10$

$\hat{\gamma}_n^{\mathrm{H}}(x)$ en ordonnée, α_n en abscisse et $\gamma(x)$ en trait interrompu



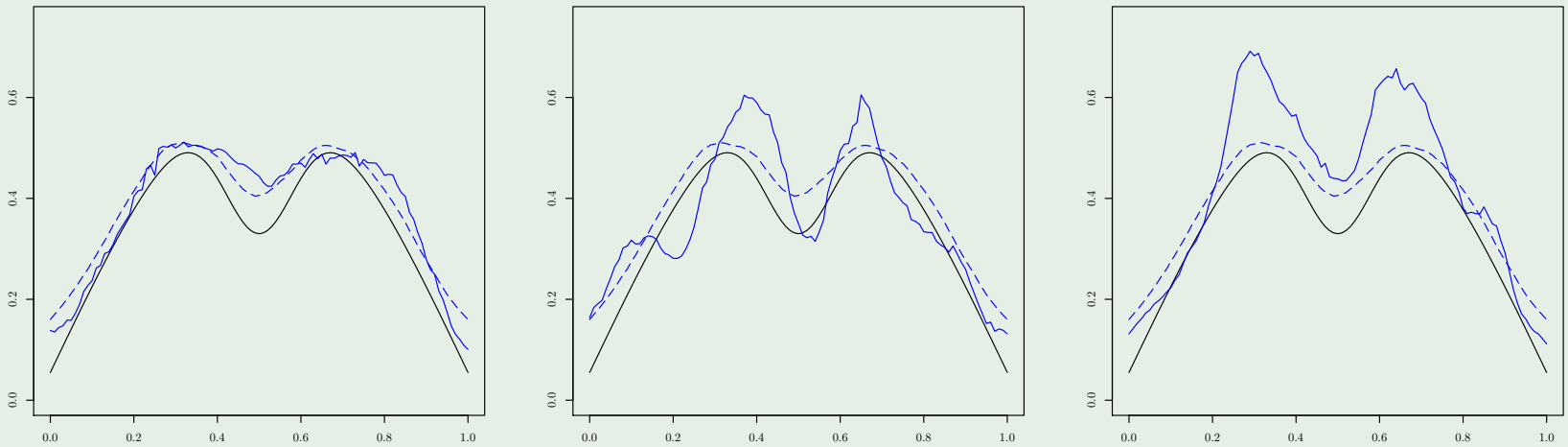
Comparaison des estimateurs de l'indice de queue $\gamma(x)$ sur une grille régulière.

— $\alpha_n = 0.250$ – Traits interrompus: les moyennes de $\hat{\gamma}_n^{\mathrm{H}}(x)$ – Gauche : 1er décile de l'erreur L_2 – Centre : médiane de l'erreur L_2 – Droite : 9ème décile de l'erreur L_2.



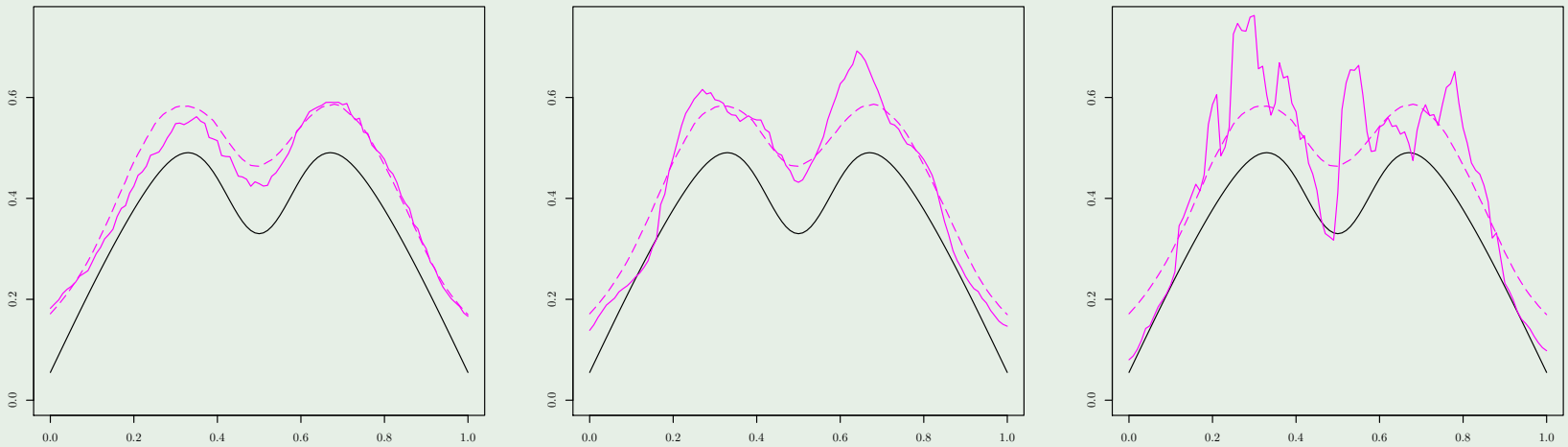
Comparaison des estimateurs de l'indice de queue $\gamma(x)$ sur une grille régulière.

— $\alpha_n = 0.50$ – Traits interrompus: les moyennes de $\hat{\gamma}_n^{\mathrm{H}}(x)$ – Gauche : 1er décile de l'erreur L_2 – Centre : médiane de l'erreur L_2 – Droite : 9ème décile de l'erreur L_2.



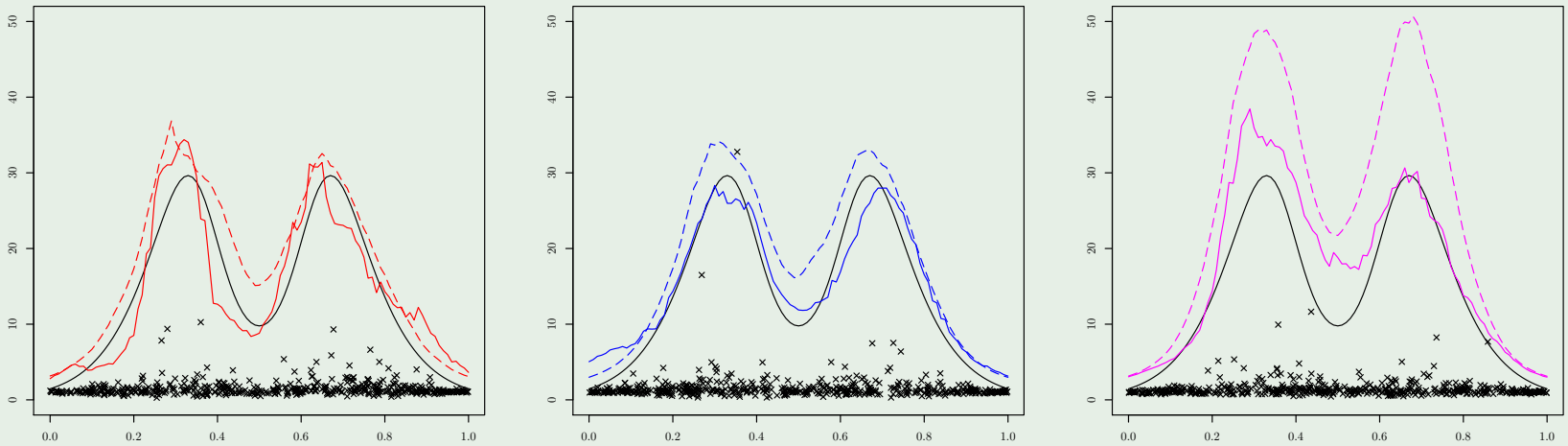
Comparaison des estimateurs de l'indice de queue $\gamma(x)$ sur une grille régulière.

— $\alpha_n = 0.75$ – Traits interrompus: les moyennes de $\hat{\gamma}_n^{\mathrm{H}}(x)$ – Gauche : 1er décile de l'erreur L_2 – Centre : médiane de l'erreur L_2 – Droite : 9ème décile de l'erreur L_2.



Comparaison du meilleur des cas de trois estimateurs de quantiles extrêmes

– $\hat{q}_n^{\mathrm{W}}(1/2n|.)$ ($\alpha_n = 0.25$) – $\hat{q}_n^{\mathrm{W}}(1/2n|.)$ ($\alpha_n = 0.50$) – $\hat{q}_n^{\mathrm{W}}(1/2n|.)$ ($\alpha_n = 0.75$) – (xxx): nuage de points – Traits interrompus : les moyennes de $\hat{q}_n^{\mathrm{W}}(1/n|.)$.



Perspectives

- Choix automatique de α_n dans la pratique.
- Extension de l'étude asymptotique de l'estimateur à noyau de la fonction de survie conditionnelle dans un contexte plus général.
- Définition d'autres estimateurs de l'indice de queue et de quantiles extrêmes conditionnels.
- Normalité asymptotique quel que soit la structure de dépendance des données (ex. cas α-mélangeant).

Bibliographie


A. Daouia, L. Gardes, S. Girard and A. Lekina. Kernel estimators of extreme level curves. http://hal.inria.fr/inria-00393588/fr/, 2010. Acceptée pour publication à TEST.


A. Berlinet, A. Gannoun and E. Matzner-Løber. Asymptotic normality of convergent estimates of conditional quantiles. *Statistics*, 35:139–169, 2001.


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


G. Collomb. *Estimation non paramétrique de la régression par la méthode du noyau*. PhD thesis, Université Paul Sabatier de Toulouse, 1976.


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


Weissman, I. Estimation of parameters and large quantiles based on the k-largest observations, *JASA*, 73, 812–815, 1978.


Q. Yao. Conditional pblueictive regions for stochastic processes. *Technical report*, University of Kent at Canterbury, 1999.