

Estimation de quantiles extrêmes pour les lois à queue de type Weibull : une synthèse bibliographique

Laurent Gardes, Stéphane Girard

▶ **To cite this version:**

Laurent Gardes, Stéphane Girard. Estimation de quantiles extrêmes pour les lois à queue de type Weibull : une synthèse bibliographique. Journal de la Societe Française de Statistique, 2013, 154 (2), pp.98-118. hal-00799249v2

HAL Id: hal-00799249
https://hal.science/hal-00799249v2

Submitted on 5 Apr 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Estimation de quantiles extrêmes pour les lois à queue de type Weibull : une synthèse bibliographique

Laurent Gardes[(1)] & Stéphane Girard[(2)]

[(1)] Université de Strasbourg & CNRS, IRMA, UMR 7501, 7 rue René Descartes,
67084 Strasbourg Cedex, France
[(2)] INRIA Rhône-Alpes, Inovallée, 655 avenue de l'Europe, Montbonnot
38334 Saint Ismier Cedex, France

abstract Cet article est une synthèse bibliographique des méthodes d'estimation de quantiles extrêmes pour les lois à queue de type Weibull. Ces lois ont une fonction de survie qui décroit vers zéro à une vitesse exponentielle. Nous montrons comment cette problématique s'inscrit plus largement dans la théorie des valeurs extrêmes.

AMS Subject Classifications: 62-00, 62G32.
Mots clés Lois à queue de type Weibull, Synthèse bibliographique.

1 Introduction

Dans cet article, nous étudions le comportement des valeurs extrêmes d'un échantillon de variables aléatoires unidimensionnelles. Nous nous concentrons essentiellement sur une famille particulière de lois : les lois à queue de type Weibull. Ces lois ont une fonction de survie qui décroit vers zéro à la vitesse exponentielle. Nous donnerons une définition plus précise de cette famille dans le paragraphe 3. Notre principal objectif est de proposer des estimateurs de quantiles extrêmes. Plus précisément, disposant d'un échantillon $X_1, \ldots, X_n$ de n variables aléatoires réelles indépendantes et identiquement distribuées de fonction de répartition commune $F(.)$, nous souhaitons estimer le réel $q(\alpha_n)$ défini par

$$q(\alpha_n) = \bar{F}^{\leftarrow}(\alpha_n), \text{ avec } \alpha_n \to 0 \text{ lorsque } n \to \infty,$$

où (α_n) est une suite connue et $\bar{F}^{\leftarrow}(u) = \inf\{x,\ \bar{F}(x) \leq u\}$ est l'inverse généralisée de la fonction de survie $\bar{F}(.) = 1 - F(.)$. Remarquons que $q(\alpha_n)$ est le quantile d'ordre $1 - \alpha_n$ de la fonction de répartition F. Un problème similaire à l'estimation de $q(\alpha_n)$ est l'estimation de "petites probabilités" p_n. Autrement dit, pour une suite de réels (x_n) fixée, nous souhaitons estimer la probabilité p_n définie par

$$p_n = \bar{F}(x_n),\ x_n \to \infty \text{ lorsque } n \to \infty.$$

Ce sont les hydrologues qui ont été parmi les premiers à s'intéresser à ces deux problèmes. Disposant d'un échantillon de hauteurs d'un cours d'eau, ils se sont posés les deux questions suivantes :

1) quelle est la hauteur d'eau qui est atteinte ou dépassée pour une faible probabilité donnée ?

2) pour une "grande" hauteur d'eau fixée, qu'elle est la probabilité d'observer une hauteur d'eau qui lui sera supérieure ?

Les questions 1) et 2) se rapportent donc respectivement à l'estimation d'un quantile extrême (ou niveau de retour en hydrologie) et d'une "petite probabilité" (ou de façon équivalente en hydrologie, période de retour). La difficulté principale réside dans le fait que l'on considère un ordre de quantile $\alpha_n \to 0$ (ou de manière équivalente un seuil $x_n \to \infty$). En effet, si par exemple $n\alpha_n \to 0$ lorsque $n \to \infty$, il est clair que

$$\mathbb{P}(X_{n,n} < q(\alpha_n)) = F^n(q(\alpha_n)) = (1 - \alpha_n)^n \to 1$$

où $X_{1,n} \leq \ldots \leq X_{n,n}$ est l'échantillon ordonné associé à $X_1, \ldots, X_n$. La quantité $q(\alpha_n)$ n'appartient donc pas à l'intervalle de variation de nos observations. En conséquence, l'estimateur de $q(\alpha_n)$ ne peut être obtenu en inversant simplement la fonction de répartition empirique

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{X_i \leq x\},$$

car $\hat{F}_n(x) = 1$ pour $x \geq X_{n,n}$. L'estimation de quantiles extrêmes et/ou de "petites probabilités" est requise dans de nombreux domaines d'application parmi lesquels citons la fiabilité [19], la finance [21], les assurances [7, 11] et la climatologie [46]. Pour répondre à ces deux questions, nous devons donc étudier de près le comportement de la queue de distribution de $F(.)$ en utilisant la théorie des valeurs extrêmes. Nous présentons les éléments essentiels de cette théorie paragraphe 2. Dans le paragraphe 3, nous proposons des estimateurs de quantiles extrêmes pour la famille des lois à queue de type Weibull.

2 Introduction à la théorie des valeurs extrêmes

Lorsque l'on s'intéresse à la partie centrale d'un échantillon, le résultat clé est le Théorème de la Limite Centrale (abrégé en TLC) donnant la loi asymptotique de la somme des observations. Par contre, si l'on souhaite étudier les valeurs extrêmes de cet échantillon, le TLC ne présente que peu d'intérêt. On utilise plutôt un résultat établissant la loi asymptotique du maximum de l'échantillon. Ce résultat est énoncé dans le paragraphe 2.1. Il permet de classer la plupart des lois en trois domaines d'attraction. La caractérisation des fonctions de répartition dans chacun de ces domaines est donnée dans le paragraphe 2.2.

2.1 Convergence en loi du maximum d'un échantillon

Le résultat ci-dessous établit la loi asymptotique du maximum $X_{n,n} = \max(X_1, \ldots, X_n)$ de l'échantillon. Il a été démontré notamment par Gnedenko [32].

Théorème 1. *S'il existe deux suites réelles $(a_n > 0)$, (b_n) et un réel γ tels que*

$$\mathbb{P}\left\{\frac{X_{n,n} - b_n}{a_n} \leq x\right\} \to H_\gamma(x),$$

lorsque $n \to \infty$ alors

$$H_\gamma(x) = \begin{cases} \exp[-(1+\gamma x)_+^{-1/\gamma}] & \text{si } \gamma \neq 0, \\ \exp(-\mathrm{e}^{-x}) \text{ si } \gamma = 0, \end{cases}$$

où $y_+ = \max(0, y)$.

La fonction de répartition $H_\gamma(.)$ est la fonction de répartition de la loi des valeurs extrêmes. La densité associée est représentée Figure 1 pour différentes valeurs de γ. Cette loi dépend du seul paramètre γ appelé l'indice des valeurs extrêmes. Selon le signe de γ, on définit trois domaines d'attraction :

- si $\gamma > 0$, on dit que $F(.)$ appartient au domaine d'attraction de Fréchet. Il contient les lois dont la fonction de survie décroit comme une fonction puissance. On parle aussi de lois à queue lourde. Dans ce domaine d'attraction, on trouve les lois de Pareto, de Student, de Cauchy, etc ...
- si $\gamma = 0$, $F(.)$ est dans le domaine d'attraction de Gumbel qui regroupe les lois ayant une fonction de survie à décroissance exponentielle. C'est le cas des lois normale, gamma, exponentielle, etc ...
- si $\gamma < 0$, $F(.)$ appartient au domaine d'attraction de Weibull. Ce domaine contient les lois dont le point terminal $x_F = \inf\{x,\ F(x) \geq 1\}$ est fini. C'est le cas par exemple des lois uniformes, lois beta, etc ...

Un classement de nombreuses lois par domaine d'attraction est disponible dans [21, Tableaux 3.4.2-3.4.4]. Un extrait en est présenté Table 1. Nous allons à présent rappeler les théorèmes de caractérisation des trois domaines d'attraction ci-dessus.

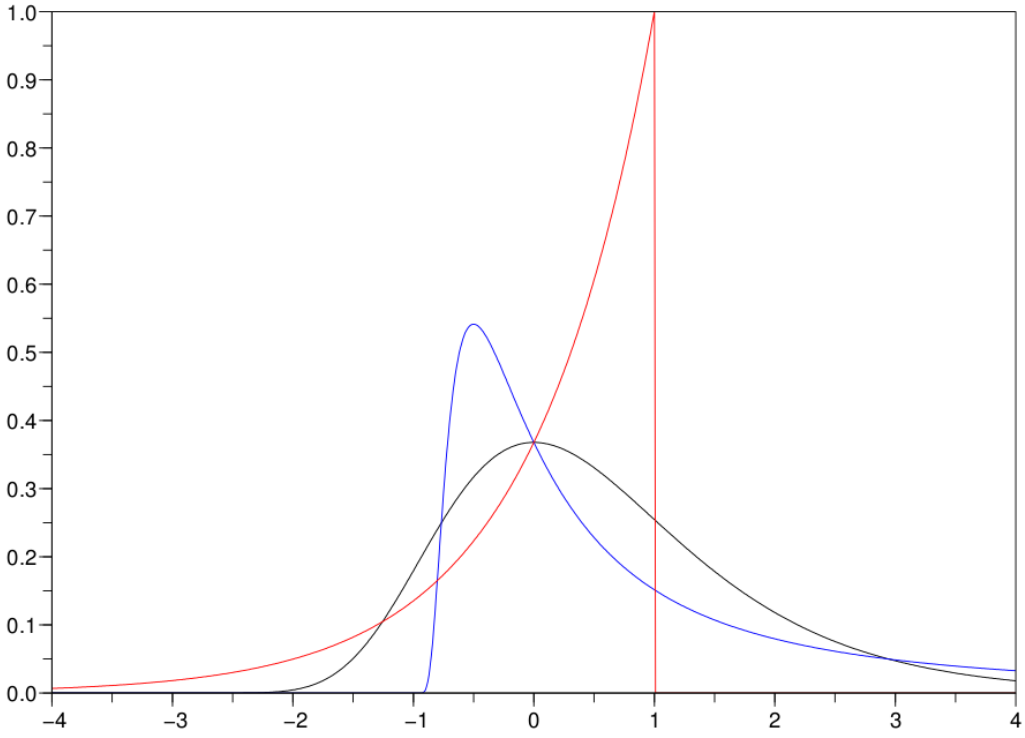


Figure 1: Exemples de densités associées à la loi des valeurs extrêmes (noir : $\gamma = 0$, bleu : $\gamma = 1$ et rouge : $\gamma = -1$).

Domaine d'attraction	**Gumbel** $\gamma = 0$	**Fréchet** $\gamma > 0$	**Weibull** $\gamma < 0$
Loi	Normale Exponentielle Lognormale Gamma Weibull	Cauchy Pareto Student Burr	Uniforme Beta

Table 1: Domaines d'attraction des lois usuelles.

2.2 Caractérisation des domaines d'attraction

La caractérisation des domaines d'attraction fait largement appel à la notion de fonctions à variations régulières. Rappelons qu'une fonction $U(.)$ positive est à variations régulières d'indice $\delta \in \mathbb{R}$ à l'infini (on notera par la suite $U(.) \in \mathcal{RV}_\delta$) si pour tout $\lambda > 0$,

$$\lim_{x\to\infty} \frac{U(\lambda x)}{U(x)} = \lambda^\delta.$$

Si $\delta = 0$, on dit que la fonction $U(.)$ est à variations lentes ($U(.) \in \mathcal{RV}_0$). On montre facilement que toute fonction à variations régulières d'indice $\delta \in \mathbb{R}$ s'écrit,

$$U(x) = x^\delta L(x),\ L \in \mathcal{RV}_0.$$

Rappelons enfin que toutes les fonctions à variations lentes $L(.)$ s'écrivent sous la forme :

$$L(x) = c(x) \exp\left\{\int_1^x \frac{\Delta(u)}{u} du\right\},$$

où $c(x) \to c > 0$ et $\Delta(x) \to 0$ lorsque $x \to \infty$. Cette représentation des fonctions à variations lentes est connue sous le nom de *représentation de Karamata* (voir [10, Théorème 1.3.1]). De plus, si la fonction $c(.)$ est constante, la fonction $L(.)$ est dite normalisée. De nombreux résultats sur les fonctions à variations régulières sont donnés dans le livre de Bingham *et al.* [10].

2.2.1 Domaine d'attraction de Fréchet

Le résultat ci-dessous énoncé par Gnedenko [32] et dont on trouvera une démonstration simple dans le livre de Resnick [45, Proposition 1.11] assure que toute fonction appartenant au domaine d'attraction de Fréchet est une fonction à variations régulières.

Théorème 2. *Une fonction de répartition $F(.)$ appartient au domaine d'attraction de Fréchet (avec un indice des valeurs extrêmes $\gamma > 0$) si et seulement si la fonction de survie $\bar{F}(.) \in \mathcal{RV}_{-1/\gamma}$.*

Autrement dit, une fonction de répartition $F(.)$ appartenant au domaine d'attraction de Fréchet s'écrit sous la forme :

$$F(x) = 1 - x^{-1/\gamma} L(x),\ L(.) \in \mathcal{RV}_0. \tag{1}$$

Les suites de normalisation (a_n) et (b_n) sont données dans ce cas par $a_n = \bar{F}^{\leftarrow}(1/n)$ et $b_n = 0$ (voir [45, Proposition 1.11]). Il faut aussi noter que toutes les fonctions de répartition du domaine d'attraction de Fréchet ont un point terminal infini. On peut montrer (voir [10, Théorème 1.5.12]) que l'équation (1) est équivalente à :

$$q(\alpha) = \alpha^{-\gamma} \ell(\alpha^{-1}),\ \ell(.) \in \mathcal{RV}_0, \tag{2}$$

où $\alpha \in [0, 1]$. De nombreux auteurs se sont intéressés à l'estimation de l'indice des valeurs extrêmes γ et des quantiles extrêmes $q(\alpha_n)$ pour des lois à queue lourde. L'estimateur le plus connu de $\gamma > 0$ est l'estimateur proposé par Hill [38] et défini par

$$\hat{\gamma}_n^H = \frac{1}{k_n} \sum_{i=1}^{k_n} \log(X_{n-i+1,n}) - \log(X_{n-k_n,n}), \tag{3}$$

où (k_n) est une suite d'entiers telle que $1 < k_n < n$. D'autres estimateurs de cet indice ont été proposés notamment par Beirlant *et al.* [6, 5] qui utilisent un modèle de régression exponentiel pour débiaiser l'estimateur de Hill et par Feuerverger *et al.* [24] qui introduisent un estimateur des moindres carrés. L'utilisation d'un noyau dans l'estimateur de Hill a été étudiée par Csörgő *et al.* [13]. Un estimateur efficace de l'indice des valeurs extrêmes a été proposé par Falk *et al.* [23]. Une liste plus détaillée des différents travaux sur l'estimation de l'indice des valeurs extrêmes est effectuée par Csörgő *et al.* [14]. Concernant l'étude du quantile extrême $q(\alpha_n)$, l'estimateur le plus fréquemment utilisé a été proposé par Weissman [48]. Il est défini par :

$$\hat{q}_n^W(\alpha_n) = X_{n-k_n+1,n} \left(\frac{k_n}{n\alpha_n}\right)^{\hat{\gamma}_n^H}. \tag{4}$$

On peut trouver de nombreux autres estimateurs du quantile extrême dans le livre écrit par de Haan et Ferreira [15].

2.2.2 Domaine d'attraction de Weibull

Le résultat suivant (voir Gnedenko [32], Resnick [45, Proposition 1.13]) montre que l'on passe du domaine d'attraction de Fréchet à celui de Weibull par un simple changement de variable dans la fonction de répartition.

Théorème 3. *Une fonction de répartition $F(.)$ appartient au domaine d'attraction de Weibull (avec un indice des valeurs extrêmes $\gamma < 0$) si et seulement si son point terminal x_F est fini et si la fonction de répartition $F_*(.)$ définie par*

$$F_*(x) = \begin{cases} 0 & \text{si } x < 0 \\ F(x_F - 1/x) & \text{si } x \geq 0, \end{cases}$$

appartient au domaine d'attraction de Fréchet avec un indice des valeurs extrêmes $-\gamma > 0$.

Ainsi, une fonction de répartition $F(.)$ du domaine d'attraction de Weibull s'écrit pour $x \leq x_F$:

$$F(x) = 1 - (x_F - x)^{-1/\gamma} L((x_F - x)^{-1}),\ L(.) \in \mathcal{RV}_0. \tag{5}$$

De manière équivalente, le quantile $q(\alpha)$ associé s'écrit :

$$q(\alpha) = x_F - \alpha^{-\gamma} \ell(1/x),\ \ell(.) \in \mathcal{RV}_0. \tag{6}$$

Les suites de normalisation (a_n) et (b_n) sont données par $a_n = x_F - \bar{F}^{\leftarrow}(1/n)$ et $b_n = x_F$. Ce domaine d'attraction a été considéré notamment par Falk [22], Girard *et al.* [31] et Hall *et al.* [37] pour estimer le point terminal d'une distribution.

2.2.3 Domaine d'attraction de Gumbel

La caractérisation des fonctions de répartition du domaine d'attraction de Gumbel est plus complexe. Le résultat ci-dessous est démontré notamment dans Resnick [45, Proposition 1.4].

Théorème 4. *Une fonction de répartition $F(.)$ appartient au domaine d'attraction de Gumbel si et seulement si il existe $z < x_F \leq \infty$ tel que*

$$\bar{F}(x) = c(x) \exp\left\{ - \int_z^x \frac{1}{a(t)} dt \right\},\ z < x < x_F, \tag{7}$$

où $c(x) \to c > 0$ lorsque $x \to x_F$ et $a(.)$ est une fonction positive et dérivable de dérivée $a'(.)$ telle que $a'(x) \to 0$ lorsque $x \to x_F$.

Le domaine d'attraction de Gumbel regroupe une grande diversité de lois comptant parmi elles la plupart des lois usuelles (loi normale, exponentielle, gamma, log-normale). Cette famille étant difficile à étudier dans toute sa généralité, de nombreux auteurs se sont concentrés sur une sous-famille : les lois à queue de type Weibull. Leur définition est donnée dans le paragraphe suivant.

3 Inférence pour les lois à queue de type Weibull

Les lois à queue de type Weibull correspondent au cas particulier où l'on suppose dans (7) que la dérivée de la fonction $a(.)$ est à variations régulières avec un indice strictement négatif. Plus précisément, si on suppose que $a'(.) \in \mathcal{RV}_{-1/\theta}$ où $\theta > 0$ est appelé l'indice de queue de Weibull, on montre facilement que l'équation (7) s'écrit :

$$\bar{F}(x) = \exp\left\{ -x^{1/\theta} L(x) \right\},\ L(.) \in \mathcal{RV}_0. \tag{8}$$

Une fonction de répartition s'écrivant selon le modèle (8) est dite à queue de type Weibull d'indice $\theta > 0$. L'équation (8) est équivalente à

$$q(\alpha) = (-\log \alpha)^{\theta} \ell(-\log \alpha),\ \ell(.) \in \mathcal{RV}_0, \tag{9}$$

où $\alpha \in [0, 1]$. Cette famille de lois contient par exemple les lois normale, Gamma, exponentielle, etc ... Par contre, la loi log-normale qui appartient au domaine d'attraction de Gumbel n'est pas une loi à queue de type Weibull. Dans la suite, on considère un échantillon $X_1, \ldots, X_n$ de variables aléatoires indépendantes et distribuées selon le modèle (8). Le paragraphe 3.1 est consacré à l'estimation de l'indice de queue de Weibull. L'estimation des quantiles extrêmes est discutée dans le paragraphe 3.2.

3.1 Estimation de l'indice de queue de Weibull

Les lois à queue de type Weibull appartiennent évidemment au domaine d'attraction de Gumbel (*i.e.* avec un indice des valeurs extrêmes $\gamma = 0$). L'indice des valeurs extrêmes ne fournit donc aucune information sur la vitesse de décroissance de la fonction de survie à l'intérieur de cette famille de loi. C'est l'indice de queue de Weibull θ qui nous donne cette information : une valeur de θ proche de zéro (resp. l'infini) correspond à une décroissance rapide (resp. lente) de la queue de distribution. La connaissance de ce paramètre est donc essentielle si l'on souhaite par exemple estimer un quantile extrême. Il existe dans la littérature de nombreux estimateurs de l'indice θ. Berred [9] propose un estimateur basé sur des valeurs records mais la majorité des estimateurs utilise les k_n plus grandes observations de l'échantillon. Parmi ceux-ci citons Beirlant *et al.* [3, 4, 8], Broniatowski [12], Diebolt *et al.* [16], Dierckx *et al.* [18], Gardes *et al.* [26, 27], Girard [30] et Goegebeur *et al.* [33, 34]. Le plus simple d'entre eux a été proposé dans [8]. Il est défini par :

$$\hat{\theta}_n^B = \sum_{i=1}^{k_n-1} (\log(X_{n-i+1,n}) - \log(X_{n-k_n+1,n})) \Bigg/ \sum_{i=1}^{k_n-1} (\log\log(n/i) - \log\log(n/k_n)) \,, \tag{10}$$

et rappelons que (k_n) est une suite d'entiers tels que $1 < k_n < n$. Son expression est proche de celle de l'estimateur de l'indice des valeurs extrêmes proposé par Hill [38] (voir l'équation (3)). Elle est basée sur la remarque suivante : d'après (9),

$$\frac{\log q(\alpha)}{\log\log(1/\alpha)} = \theta + \frac{\log \ell(\log(1/\alpha))}{\log\log(1/\alpha)}.$$

Ainsi, comme $\log(\ell(x))/\log(x) \to 0$ lorsque $x \to \infty$ (voir [10, Propriété 1.3.6]), on en déduit que pour $\alpha \to 0$,

$$\log q(\alpha) \sim \theta \log\log(1/\alpha). \tag{11}$$

Ainsi, les points $(\log\log(n/i), \log(X_{n-i+1,n}))$, $i = 1, \ldots, k_n - 1$ sont approximativement répartis sur une droite de pente θ. Nous détaillons dans les paragraphes 3.1.1 et 3.1.2 deux familles d'estimateurs englobant $\hat{\theta}_n^B$ ainsi qu'un estimateur débiaisé de l'indice θ (voir le paragraphe 3.1.3). Les résultats asymptotiques sont obtenus (entre autres) sous les hypothèses suivantes.

(H.1) La suite (k_n) vérifie $k_n \to \infty$ et $n/k_n \to \infty$ lorsque $n \to \infty$.

(H.2) Il existe un paramètre $\rho < 0$ et une fonction $b(.)$ vérifiant $b(x) \to 0$ lorsque $x \to \infty$ tels que pour tout $1 < A < \infty$

$$\lim_{x\to\infty} \sup_{\lambda\in[1,A]} \left| \frac{\log(\ell(\lambda x)/\ell(x))}{b(x)K_\rho(\lambda)} - 1 \right| = 0,$$

où $K_\rho(\lambda) = \int_1^\lambda t^{\rho-1}dt$ et $\ell(.)$ est la fonction à variations lentes introduite dans (9).

L'hypothèse **(H.1)** assure que le nombre de statistiques d'ordre conservées k_n est assez grand ($k_n \to \infty$) pour obtenir des estimateurs stables, mais pas trop ($n/k_n \to \infty$) pour que les observations utilisées restent dans la queue de distribution. Le choix de la suite (k_n) est donc un compromis entre le biais et la variance de l'estimateur.
L'hypothèse **(H.2)** est très souvent utilisée pour étudier le comportement asymptotique d'estimateurs d'indice ou de quantiles extrêmes. Elle est notamment nécessaire pour démontrer la normalité asymptotique de l'estimateur de Hill défini en (3). On peut montrer (voir par exemple [29]) que la fonction $b(.)$ (appelée aussi fonction de biais) est à variations régulières d'indice $\rho < 0$. Le paramètre ρ (appelé paramètre du second ordre) contrôle donc la vitesse de convergence de $\ell(\lambda x)/\ell(x)$ vers 1. Une valeur de ρ proche de 0 implique une faible vitesse de convergence. Quelques exemples en sont donnés dans la Table 2.

Loi	θ	$b(x)$	ρ
Normale $\mathcal{N}(\mu, \sigma^2)$	$1/2$	$\frac{1}{4}\frac{\log x}{x}$	-1
Gamma $\Gamma(\alpha \neq 1, \lambda)$	1	$(1-\alpha)\frac{\log x}{x}$	-1
Weibull $\mathcal{W}(\alpha, \lambda)$	$1/\alpha$	0	$-\infty$

Table 2: Paramètres θ, ρ et fonction $b(.)$ associés aux lois usuelles. Les paramètres α et λ sont respectivement des paramètres de forme et d'échelle.

3.1.1 Utilisation de poids

Nous nous focalisons sur la famille d'estimateurs [27] de θ incorporant des poids dans l'estimateur $\hat{\theta}_n^B$ défini par (10). Les estimateurs obtenus sont donc des combinaisons linéaires de statistiques d'ordre c'est-à-dire des L-estimateurs. Plus précisément, on considère la famille d'estimateurs $\Theta_1 = \{\hat{\theta}_n(\zeta),\ \zeta = (\zeta_{1,n}, \ldots, \zeta_{k_n-1,n})\}$ avec

$$\hat{\theta}_n(\zeta) = \sum_{i=1}^{k_n-1} \zeta_{i,n}(\log(X_{n-i+1,n}) - \log(X_{n-k_n+1,n})) \Big/ \sum_{i=1}^{k_n-1} \zeta_{i,n}(\log\log(n/i) - \log\log(n/k_n)) \ , \quad (12)$$

où $\zeta_{i,n} = W(i/k_n) + \varepsilon_{i,n}$, $(\varepsilon_{i,n})$, $i = 1, \ldots, k_n - 1$ étant une suite non-aléatoire. La fonction déterministe $W(.)$ doit satisfaire les deux hypothèses de régularité ci dessous :

(H.3) la fonction $W(.)$ est définie et admet une dérivée continue sur l'intervalle $]0, 1[$.

(H.4) Il existe $M > 0$, $0 \leq q < 1/2$ et $p < 1$ tels que pour tout $x \in]0, 1[$, $|W(x)| \leq Mx^{-q}$ et $|W'(x)| \leq Mx^{-p-q}$.

Ces conditions sont essentielles pour établir la normalité asymptotique des L-estimateurs (voir par exemple [41]). On pose :

$$\|\varepsilon\|_{n,\infty} = \max_{1,\ldots,k_n-1} |\varepsilon_{i,n}|,\ \mu(W) = \int_0^1 W(x)\log(1/x)dx,$$

$$\sigma^2(W) = \int_0^1 \int_0^1 W(x)W(y)\frac{\min(x,y) - xy}{xy}dxdy.$$

Nous rappelons à présent (voir [27, Théorème 1]) le résultat de normalité asymptotique des estimateurs appartenant à cette famille.

Théorème 5. *On se place sous le modèle (8) et on suppose que les conditions* **(H.1)** *à* **(H.4)** *sont satisfaites. Si* $k_n^{1/2} b(\log(n/k_n)) \to \Lambda \in \mathbb{R}$ *et* $k_n^{1/2} \max\{1/\log(n), \|\varepsilon\|_{n,\infty}\} \to 0$ *lorsque* $n \to \infty$ *alors,*

$$k_n^{1/2}(\hat{\theta}_n(\zeta) - \theta - b(\log(n/k_n))) \xrightarrow{d} \mathcal{N}(0, \theta^2\sigma^2(W)/\mu^2(W)).$$

Dans le cas où $\liminf \|\varepsilon\|_{n,\infty} \log(n) \leq 1$ avec $\Lambda \neq 0$, le Théorème 5 n'est valable que si $\rho > -1$ ce qui correspond à une vitesse de convergence lente dans la condition **(H.2)**. Donnons à présent deux choix

possibles pour les poids $\zeta_{i,n}$, $i = 1, \ldots, k_n - 1$.

En prenant $\zeta_{i,n} = 1$ pour tout $i = 1, \ldots, k_n - 1$ (*i.e.* $W(x) = 1$ pour tout $x \in]0, 1[$ et $\varepsilon_{i,n} = 0$ pour tout $i = 1, \ldots, k_n - 1$), on retrouve l'estimateur $\hat{\theta}_n^B$ proposé par Beirlant *et al.* [8] (voir l'équation (10)). Le résultat de normalité asymptotique établi par Girard [30] est une conséquence directe du Théorème 5. Nous rappelons ce résultat ci-dessous :

Corollaire 1. *On se place sous le modèle (8) et on suppose que les conditions* **(H.1)** *et* **(H.2)** *sont satisfaites. Si* $k_n^{1/2} b(\log(n/k_n)) \to 0$ *et* $k_n^{1/2}/\log(n) \to 0$ *alors* $k_n^{1/2}(\hat{\theta}_n^B - \theta) \xrightarrow{d} \mathcal{N}(0, \theta^2)$.

En rappelant que les points $(\log\log(n/i), \log(X_{n-i+1,n}))$, $i = 1, \ldots, k_n - 1$ sont approximativement répartis sur une droite de pente θ, il est possible d'estimer θ par l'estimateur des moindres carrés. Il est montré dans [27, Corollaire 2] que l'estimateur des moindres carrés appartient à la famille Θ_1 avec les poids :

$$\zeta_{i,n} = \zeta_{i,n}^Z = \log\log(n/i) - \frac{1}{k_n - 1} \sum_{i=1}^{k_n - 1} \log\log(n/i) = W(i/k_n) + \varepsilon_{i,n},$$

où $W(x) = -(\log(x) + 1)$ et, uniformément en $i = 1, \ldots, k_n - 1$, $\varepsilon_{i,n} = \mathrm{O}(\log^2(k_n)/\log(n)) + \mathrm{O}(\log(k_n)/k_n)$. L'estimateur $\hat{\theta}_n(\zeta^Z)$ ainsi obtenu est similaire à l'estimateur de l'indice des valeurs extrêmes de Zipf introduit par Kratz *et al.* [39] et Schultze *et al.* [47] dans le cas de lois à queue lourde. Nous rappelons ci-dessous le résultat sur la normalité asymptotique de $\hat{\theta}_n(\zeta^Z)$ (voir [27, Corollaire 2]). C'est une conséquence directe du Théorème 5.

Corollaire 2. *On se place sous le modèle (8) et on suppose que les conditions* **(H.1)** *et* **(H.2)** *sont satisfaites. Si* $k_n^{1/2} b(\log(n/k_n)) \to 0$ *et* $k_n^{1/2} \log^2(k_n)/\log(n) \to 0$ *alors* $k_n^{1/2}(\hat{\theta}_n(\zeta^Z) - \theta) \xrightarrow{d} \mathcal{N}(0, 2\theta^2)$.

3.1.2 Utilisation d'autres suites de normalisation

Dans l'estimateur $\hat{\theta}_n^B$ (voir l'équation (10)), la somme des écarts entre les logarithmes des statistiques d'ordre est normalisée par la suite :

$$T_n^{(1)} = \sum_{i=1}^{k_n - 1} (\log\log(n/i) - \log\log(n/k_n)). \tag{13}$$

Il est possible de remplacer cette suite $(T_n^{(1)})$ par une suite positive quelconque (T_n). Ceci conduit à définir la famille d'estimateurs $\Theta_2 = \{\hat{\theta}_n(T_n),\ T_n > 0\}$ avec

$$\hat{\theta}_n(T_n) = \frac{1}{T_n} \sum_{i=1}^{k_n - 1} (\log(X_{n-i+1,n}) - \log(X_{n-k_n+1,n})). \tag{14}$$

En posant

$$u_n = \frac{k_n \int_0^\infty \log(1 + x/t) \mathrm{e}^{-x} dx}{T_n} - 1,$$

on obtient [26, Théorème 2.1] un résultat de normalité asymptotique pour cette famille d'estimateurs. Ce résultat est rappelé ci-dessous :

Théorème 6. *On se place sous le modèle (8) et on suppose que les conditions* **(H.1)** *et* **(H.2)** *sont satisfaites. Si* $T_n k_n \log(n/k_n) \to 1$ *et* $k_n^{1/2} b(\log(n/k_n)) \to \Lambda \in \mathbb{R}$ *lorsque* $n \to \infty$ *alors*

$$k_n^{1/2}(\hat{\theta}_n(T_n) - \theta - b(\log(n/k_n)) - \theta u_n) \xrightarrow{d} \mathcal{N}(0, \theta^2).$$

La meilleure vitesse de convergence de $\hat{\theta}_n(T_n)$ est obtenue lorsque $\Lambda \neq 0$. Dans ce cas, on peut montrer (voir [26, Proposition 2.2]) que (k_n) est équivalente à $\Lambda^2 (\log n)^{-2\rho} \ell^*(\log(n))$ où $\ell^*(.)$ est une fonction à variations lentes. Sous l'hypothèse supplémentaire $\liminf \|\varepsilon\|_{n,\infty} \log(n) \leq 1$, les estimateurs de la

famille Θ_1 ont la même vitesse de convergence que ceux de la famille Θ_2. Le Théorème 6 permet de déterminer l'erreur moyenne quadratique asymptotique (AMSE) de $\hat{\theta}_n(T_n)$. Elle est donnée par

$$AMSE(\hat{\theta}_n(T_n)) = (\theta u_n + b(\log(n/k_n)))^2 + \frac{\theta^2}{k_n}. \tag{15}$$

Le terme de variance asymptotique est donc identique pour tous les estimateurs de la famille. Le biais dépend par contre du choix de la suite (T_n). Le choix idéal pour cette suite de normalisation serait donc de prendre T_n de telle sorte que $\theta u_n + b(\log(n/k_n)) = 0$. Malheureusement, θ et la fonction de biais $b(.)$ sont inconnus et il est donc impossible de définir une telle suite (T_n). Donnons à présent quelques choix possibles pour cette suite.

Comme nous l'avons déjà mentionné, le choix $(T_n) = (T_n^{(1)})$ (voir l'équation (13)) conduit à l'estimateur $\hat{\theta}_n^B$ (voir l'équation (10)). La normalité asymptotique de $\hat{\theta}_n(T_n^{(1)}) = \hat{\theta}_n^B$ énoncée ci-dessous est une conséquence directe du Théorème 6.

Corollaire 3. *On se place sous le modèle (8) et on suppose que les conditions* **(H.1)** *et* **(H.2)** *sont satisfaites. Si* $k_n^{1/2} b(\log(n/k_n)) \to 0$ *et* $\log(k_n)/\log(n) \to 0$ *alors* $k_n^{1/2}(\hat{\theta}_n(T_n^{(1)}) - \theta) \xrightarrow{d} \mathcal{N}(0, \theta^2)$.

Dans le Corollaire 1 du sous-paragraphe précédent, ce résultat est obtenu sous la condition supplémentaire $k_n^{1/2}/\log(n) \to 0$.

Un choix naturel pour (T_n) est de prendre la suite annulant une partie du biais asymptotique : la partie dépendant de u_n. Pour ce faire, il suffit de prendre $(T_n) = (T_n^{(2)})$ avec

$$T_n^{(2)} = k_n \int_0^\infty \log\left(1 + \frac{x}{\log(n/k_n)}\right) \mathrm{e}^{-x} dx.$$

En remarquant que $T_n^{(2)} = n/k_n \mathrm{E}_1(\log(n/k_n))$, où $\mathrm{E}_1(z)$ dénote l'exponentielle intégrale calculée au point z (voir [1, Chapitre 5, p. 225-233]), le calcul de la suite $T_n^{(2)}$ est simple à effectuer. Il est aussi intéressant de noter que

$$T_n^{(1)} = \sum_{i=1}^{k_n} \log\left(1 - \frac{\log(i/k_n)}{\log(n/k_n)}\right)$$

est en fait l'approximation par des sommes de Riemann de $T_n^{(2)}$ car en intégrant par partie on montre que

$$T_n^{(2)} = \int_0^1 \log\left(1 - \frac{\log(x)}{\log(n/k_n)}\right) dx.$$

On déduit du Théorème 6 le résultat suivant :

Corollaire 4. *On se place sous le modèle (8) et on suppose que les conditions* **(H.1)** *et* **(H.2)** *sont satisfaites. Si* $k_n^{1/2} b(\log(n/k_n)) \to 0$ *alors* $k_n^{1/2}(\hat{\theta}_n(T_n^{(2)}) - \theta) \xrightarrow{d} \mathcal{N}(0, \theta^2)$.

Une des hypothèses du Théorème 6 étant que $T_n k_n \log(n/k_n) \to 1$, un choix simple est de prendre $T_n = T_n^{(3)} = (k_n \log(n/k_n))^{-1}$. La normalité asymptotique de l'estimateur ainsi obtenu est donnée ci-dessous :

Corollaire 5. *On se place sous le modèle (8) et on suppose que les conditions* **(H.1)** *et* **(H.2)** *sont satisfaites. Si* $k_n^{1/2} b(\log(n/k_n)) \to 0$ *et* $k_n^{1/2}/\log(n/k_n) \to 0$ *alors* $k_n^{1/2}(\hat{\theta}_n(T_n^{(3)}) - \theta) \xrightarrow{d} \mathcal{N}(0, \theta^2)$.

Notez que les Corollaires 3, 4 et 5 sont énoncés dans [26, Corollaire 1] et sont démontrés dans ce même article. Nous allons à présent comparer ces trois estimateurs en fonction de leur erreur moyenne quadratique asymptotique définie par l'équation (15).

Proposition 1. *On se place sous le modèle (8) et on suppose que les conditions* **(H.1)** *et* **(H.2)** *sont satisfaites. Si* $k_n^{1/2} b(\log(n/k_n)) \to \Lambda \in \mathbb{R}$*, plusieurs situations sont possibles. On pose :*

$$\beta_1 = 2 \lim_{x\to\infty} x b(x) \text{ et } \beta_2 = -4 \lim_{n\to\infty} b(\log n) \frac{k_n}{\log k_n}.$$

i) Soit $b(.)$ est asymptotiquement strictement positive.

Si $\beta_1 > \theta$ alors, pour n assez grand,

$$AMSE(\hat{\theta}_n(T_n^{(3)})) < AMSE(\hat{\theta}_n(T_n^{(2)})) < AMSE(\hat{\theta}_n(T_n^{(1)})).$$

Si $\beta_1 < \theta$ alors, pour n assez grand,

$$AMSE(\hat{\theta}_n(T_n^{(2)})) < \min(AMSE(\hat{\theta}_n(T_n^{(1)})), AMSE(\hat{\theta}_n(T_n^{(3)}))).$$

ii) Soit $b(.)$ est asymptotiquement strictement négative.

Si $\beta_2 > \theta$, alors, pour n assez grand,

$$AMSE(\hat{\theta}_n(T_n^{(1)})) < AMSE(\hat{\theta}_n(T_n^{(2)})) < AMSE(\hat{\theta}_n(T_n^{(3)})).$$

Si $\beta_2 < \theta$, alors, pour n assez grand,

$$AMSE(\hat{\theta}_n(T_n^{(2)})) < \min(AMSE(\hat{\theta}_n(T_n^{(1)})), AMSE(\hat{\theta}_n(T_n^{(3)}))).$$

Cette proposition a été établie dans [26, Théorème 3]. Comme l'on pouvait s'y attendre, il n'y a pas d'estimateur qui soit préférable dans toutes les situations. Dans le cas i), β_1 ne dépend pas de la suite (k_n). Le classement entre les trois estimateurs $\hat{\theta}_n(T_n^{(1)})$, $\hat{\theta}_n(T_n^{(2)})$ et $\hat{\theta}_n(T_n^{(3)})$ dépend donc uniquement de la loi des observations. Si la fonction biais $b(.)$ converge rapidement vers zéro alors $\beta_1 < \theta$ et ainsi l'utilisation de l'estimateur $\hat{\theta}_n(T_n^{(2)})$ est préférable. Au contraire, si $b(.)$ converge lentement vers zéro, $\beta_1 > \theta$ et l'estimateur $\hat{\theta}_n(T_n^{(3)})$ sera de meilleure qualité. Dans le cas ii), β_2 dépend de la suite (k_n). Si (k_n) est petite (par exemple $k_n \propto -1/b(\log(n))$) alors $\beta_2 = 0$ et l'estimateur $\hat{\theta}_n(T_n^{(2)})$ est préférable. Inversement, si (k_n) est grande (par exemple $k_n \propto (b(log(n)))^{-2}$) alors $\beta_2 = \infty$ et $\hat{\theta}_n(T_n^{(1)})$ sera asymptotiquement le meilleur estimateur. Enfin, si $\beta_1 = \theta$, on ne peut pas comparer dans la situation i) les estimateurs $\hat{\theta}_n(T_n^{(2)})$ et $\hat{\theta}_n(T_n^{(3)})$ car $AMSE(\hat{\theta}_n(T_n^{(3)})) - AMSE(\hat{\theta}_n(T_n^{(2)})) \to 0$ lorsque n tend vers l'infini. De même, dans la situation ii), on ne peut pas comparer les estimateurs $\hat{\theta}_n(T_n^{(1)})$ et $\hat{\theta}_n(T_n^{(2)})$ si $\beta_2 = \theta$. Les comparaisons ci-dessus sont uniquement valables asymptotiquement. Afin de voir si elles sont aussi valides pour une taille d'échantillon finie, nous simulons $N = 200$ échantillons de taille $n = 500$ d'une loi $\Gamma(0.5, 1)$. Pour chaque échantillon, nous calculons les estimateurs $\hat{\theta}_{n,i}(T_n^{(1)})$, $\hat{\theta}_{n,i}(T_n^{(2)})$ et $\hat{\theta}_{n,i}(T_n^{(3)})$ $(i = 1, \ldots, N)$ pour différentes valeurs de k_n. Nous en déduisons les erreurs en moyenne quadratique empiriques définies par :

$$\frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_{n,i}(T_n^{(j)}) - \theta)^2,\ j = 1,\ 2,\ 3.$$

Ces erreurs empiriques sont comparées aux erreurs asymptotiques des trois estimateurs données par l'équation (15). Les résultats sont présentés dans la Figure 2 où l'on s'aperçoit que le comportement des erreurs empiriques et celui des erreurs asymptotiques sont très similaires. Les résultats de comparaison des trois estimateurs données dans le Théorème 1 semblent donc aussi utilisables pour une taille d'échantillon finie. D'autres simulations avec différentes lois ont été faites (voir [26]) conduisant à la même conclusion.

3.1.3 Un estimateur de θ débiaisé

Considérons à présent un estimateur débiaisé de l'indice de queue de Weibull θ. Il est basé sur un modèle de régression exponentiel inspiré de ceux proposés par Beirlant *et al.* [5, 6] et Feuerverger *et al.* [24] afin d'estimer l'indice des valeurs extrêmes pour des lois du domaine d'attraction de Fréchet. Plus précisément, on définit les variables aléatoires

$$Z_j = j\log(n/j)(\log(X_{n-j+1,n}) - \log(X_{n-j,n})),\ j = 1, \ldots, k_n.$$

Le modèle suivant peut être établi [16, Corollaire 2.1] :

$$Z_j = \left(\theta + \left(\frac{\log(n/k_n)}{\log(n/j)}\right) b(\log(n/k_n))\right) f_j + o_P(b(\log(n/k_n))),\ j = 1, \ldots, k_n \tag{16}$$

où f_j, $j = 1, \ldots, k_n$ sont des variables aléatoires indépendantes de loi exponentielle de paramètre 1 et le terme $o_P(b(\log(n/k_n)))$ ne dépend pas de j. On obtient à partir du modèle (16) l'approximation

$$Z_j \approx \theta + b(\log(n/k_n))x_j + \eta_j,\ j = 1, \ldots, k_n \tag{17}$$

où η_j est un terme d'erreur aléatoire centré et $x_j = \log(n/k_n)/\log(n/j)$. En estimant les paramètres θ et $b(\log(n/k_n))$ du modèle de régression linéaire (17) par la méthode des moindres carrés ordinaires, on obtient un estimateur de θ débiaisé :

$$\hat{\theta}_n^D = \frac{1}{k_n}\sum_{j=1}^{k_n} Z_j - \frac{\hat{b}(\log(n/k_n))}{k_n}\sum_{j=1}^{k_n} x_j, \tag{18}$$

où

$$\hat{b}(\log(n/k_n)) = \sum_{j=1}^{k_n}\left(x_j - \frac{1}{k_n}\sum_{j=1}^{k_n} x_j\right) Z_j \Bigg/ \sum_{j=1}^{k_n}\left(x_j - \frac{1}{k_n}\sum_{j=1}^{k_n} x_j\right)^2. \tag{19}$$

La normalité asymptotique de $\hat{\theta}_n^D$ est donnée par le résultat ci-dessous (voir [16, Théorème 3.1]).

Théorème 7. *On se place sous le modèle (8) et on suppose que les conditions* **(H.1)** *et* **(H.2)** *sont satisfaites. Supposons de plus que la fonction $b(.)$ est telle que $x|b(x)| \to \infty$ lorsque $x \to \infty$ avec*

$$\frac{k_n^{1/2}}{\log(n/k_n)} b(\log(n/k_n)) \to \tilde{\Lambda} \in \mathbb{R}.$$

Supposons enfin que, si $\tilde{\Lambda} = 0$, $\dfrac{\log^2(k_n)}{\log(n/k_n)} \to 0$ et $\dfrac{k_n^{1/2}}{\log(n/k_n)} \to \infty$. On a :

$$\frac{k_n^{1/2}}{\log(n/k_n)}(\hat{\theta}_n^D - \theta) \xrightarrow{d} \mathcal{N}(0, \theta^2).$$

L'hypothèse $x|b(x)| \to \infty$ implique que dans la condition **(H.2)** la vitesse de convergence est lente (et plus particulièrement que $\rho \geq -1$). Les estimateurs non débiaisés de θ auront donc tendance à avoir un biais important dans ce cas. On peut montrer (voir Table 2) que les lois normale, Gamma satisfont cette hypothèse mais pas les lois de Weibull. En prenant $\Lambda \neq 0$ dans les Théorèmes 5 et 6 et $\tilde{\Lambda} \neq 0$ dans le Théorème 7, on montre que l'estimateur débiaisé $\hat{\theta}_n^D$ (voir l'équation (18)) admet la même vitesse de convergence que les estimateurs des familles Θ_1 et Θ_2 avec en plus un biais asymptotique nul.

3.1.4 Illustration sur simulations

Nous effectuons une simulation afin de comparer le comportement de l'estimateur non débiaisé $\hat{\theta}_n^B$ (voir l'équation (10)) avec celui de l'estimateur débiaisé $\hat{\theta}_n^D$ (voir l'équation (18)). Nous générons $N = 100$ échantillons de taille $n = 500$ selon une loi normale centrée et réduite (pour laquelle $\theta = 1/2$). En fonction du nombre de statistiques ordonnées k_n, nous comparons sur la Figure 3 les moyennes empiriques des deux estimateurs ainsi que leurs erreurs moyennes quadratiques empiriques. Nous notons un bon comportement de l'estimateur débiaisé en terme de biais ainsi qu'en terme de variance. La méthode de débiaisage proposée semble donc efficace à taille d'échantillon finie. Des simulations sur d'autres lois ont confirmé ce bon comportement (voir [16, Section 4]).

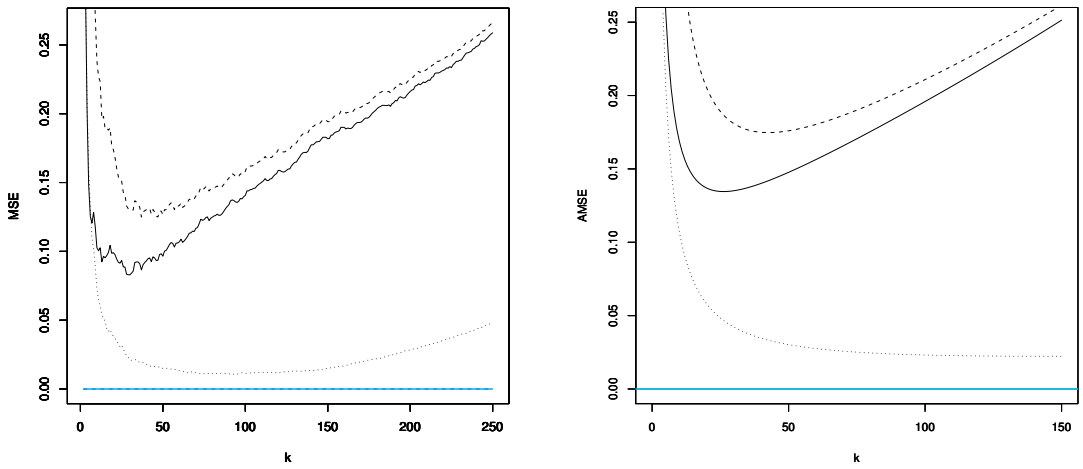


Figure 2: Comparaison des estimateurs $\hat{\theta}_n(T_n^{(1)})$ (trait plein), $\hat{\theta}_n(T_n^{(2)})$ (tirés) et $\hat{\theta}_n(T_n^{(3)})$ (pointillés). En abscisse, k_n et en ordonnée les erreurs moyenne quadratique empiriques (gauche) et asymptotiques (droite).

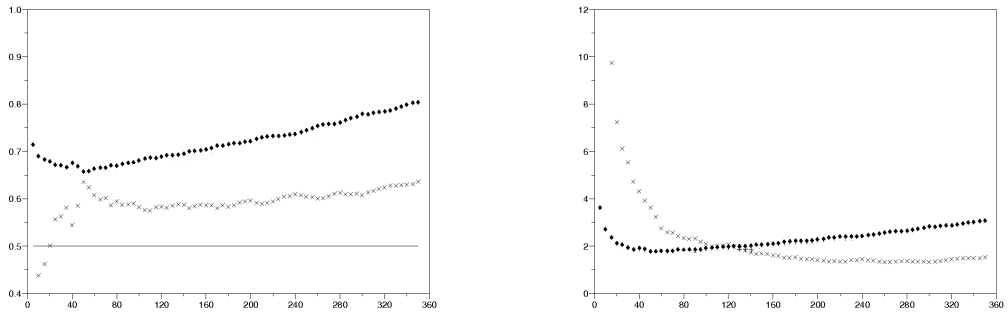


Figure 3: Comparaison des estimateurs $\hat{\theta}_n^D$ ($\times$) et $\hat{\theta}_n^B$ ($\diamond$). En abscisse, k_n et en ordonnée les erreurs en moyenne quadratique empiriques (gauche) et asymptotiques (droite).

3.1.5 Choix du nombre k_n de statistiques d'ordre

En ne tenant pas compte du terme de biais dans le modèle de régression (16), on obtient un estimateur non débiaisé de θ défini par :

$$\frac{1}{k_n}\sum_{j=1}^{k_n} Z_j.$$

L'erreur moyenne quadratique asymptotique (AMSE) de cet estimateur est donnée par [16, Théorème 2.2] :

$$AMSE(k_n) = \frac{\theta^2}{k_n} + \left(\frac{b(\log(n/k_n))}{k_n}\sum_{j=1}^{k_n}\frac{\log(n/k_n)}{\log(n/j)}\right)^2.$$

Un choix possible pour k_n est alors de prendre $k_n^{opt} = \arg\min_{k_n} AMSE(k_n)$. Nous pouvons estimer cette erreur par la quantité $\widehat{AMSE}(k_n)$ obtenue en remplaçant θ et $b(\log(n/k_n)$ par les estimateurs $\hat{\theta}_n^D$ et $\hat{b}(\log(n/k_n)$ définis précédemment (voir les équations (18) et (19)). Le nombre k_n^{opt} est estimé par :

$$\hat{k}_n = \arg\min_{k_n} \widehat{AMSE}(k_n).$$

Comme l'ont fait remarquer récemment Asimit *et al.* [2], $AMSE(k_n) \sim \theta^2/k_n + b^2(\log(n))$. Ainsi, la sélection du nombre d'observations k_n n'est pas justifiée théoriquement puisque $k_n^{opt} \sim n$. Cependant, les simulations effectuées dans [16, Section 4] montrent que l'utilisation de la valeur $\hat{k}_n$ pour estimer l'indice θ conduit à de bons résultats. Une méthode alternative a récemment été proposée [42] basée sur les idées de [20] elles-mêmes inspirées d'une variante de la méthode de Lepski [40].

3.2 Estimation de quantiles extrêmes

Toujours pour des lois à queue de type Weibull, nous nous intéressons à présent au problème d'estimation d'un quantile extrême $q(\alpha_n)$ lorsque l'ordre α_n converge vers zéro. Le principal estimateur de $q(\alpha_n)$ disponible dans la littérature a été proposé par Beirlant *et al.* [8]. Il est basé sur l'approximation (11) qui assure que pour n assez grand, on a sous l'hypothèse **(H.1)** :

$$\log q(\alpha_n) \approx \theta \log\log(1/\alpha_n) \text{ et } \log q(k_n/n) \approx \theta \log\log(n/k_n).$$

En soustrayant membre à membre les deux approximations ci-dessus et en appliquant la fonction exponentielle, on montre facilement que :

$$q(\alpha_n) \approx q(k_n/n)\left(\frac{\log(1/\alpha_n)}{\log(n/k_n)}\right)^{\theta}.$$

En estimant $q(k_n/n)$ par $X_{n-k_n+1,n}$ qui est le quantile associé à la fonction de répartition empirique et θ par $\hat{\theta}_n^B$ (voir équation (10)), Beirlant *et al.* [8] proposent l'estimateur suivant :

$$\hat{q}^B(\alpha_n) = X_{n-k_n+1,n}\left(\frac{\log(1/\alpha_n)}{\log(n/k_n)}\right)^{\hat{\theta}_n^B}.$$

La construction de cet estimateur est similaire à celle de l'estimateur proposé par Weissman [48] (voir équation (4)) pour des lois du domaine d'attraction de Fréchet. Un autre estimateur de $q(\alpha_n)$ a été proposé par Beirlant *et al.* [4]. Il est défini par :

$$\hat{q}^{B*}(\alpha_n) = X_{n-k_n+1,n}\left(1 + \frac{\hat{\sigma}_n \log(k_n/(n\alpha_n))}{\hat{\theta}_n^{B*} X_{n-k_n+1,n}}\right)^{\hat{\theta}_n^{B*}},$$

avec

$$\hat{\sigma}_n = \frac{1}{k_n - 1}\sum_{i=1}^{k_n-1}(X_{n-i+1,n} - X_{n-k_n+1,n}) \text{ et } \hat{\theta}_n^{B*} = \frac{\log n/k_n}{X_{n-k_n+1,n}}\hat{\sigma}_n.$$

Etant donné que

$$1+\frac{\hat{\sigma}_n \log(k_n/(n\alpha_n))}{\hat{\theta}_n^{B*} X_{n-k_n+1,n}} = \frac{\log(1/\alpha_n)}{\log(n/k_n)},$$

l'estimateur $\hat{q}^{B*}(\alpha_n)$ est en fait l'estimateur $\hat{q}^{B}(\alpha_n)$ pour lequel l'indice θ n'est pas estimé par $\hat{\theta}_n^B$ mais par $\hat{\theta}_n^{B*}$. L'étude du comportement asymptotique de ces deux estimateurs peut alors être unifiée [25]. Plus généralement, intéressons-nous à la famille d'estimateurs du quantile extrême $q(\alpha_n)$ définie par

$$\mathcal{Q}_{\alpha_n} = \{\hat{q}(\alpha_n, \hat{\theta}_n),\ \hat{\theta}_n \text{ estimateur de } \theta\},$$

avec

$$\hat{q}(\alpha_n, \hat{\theta}_n) = X_{n-k_n+1,n} \tau_n^{\hat{\theta}_n},\ \tau_n = \frac{\log(1/\alpha_n)}{\log(n/k_n)}, \tag{20}$$

où $\hat{\theta}_n$ est un estimateur quelconque de θ.

3.2.1 Etude de la famille d'estimateurs $\mathcal{Q}_{\alpha_n}$

Nous nous proposons d'établir la loi asymptotique des estimateurs de la famille $\mathcal{Q}_{\alpha_n}$. Deux situations peuvent se présenter : soit l'estimateur $\hat{\theta}_n$ converge rapidement vers θ de telle sorte que la loi asymptotique de $\hat{q}(\alpha_n, \hat{\theta}_n)$ est donnée par celle de la statistique d'ordre $X_{n-k_n+1,n}$. Cette situation se présente lorsque la condition ci-dessous est satisfaite :

(H.5) Il existe une suite β_n telle que $\log(n/k_n) k_n^{1/2}(\hat{\theta}_n - \beta_n - \theta) \xrightarrow{\mathrm{P}} 0$.

Soit la vitesse de convergence de $\hat{\theta}_n$ vers θ est inférieure à $\log(n/k_n)k_n^{1/2}$ et dans ce cas la loi asymptotique est celle de l'estimateur de l'indice θ. Cette situation est décrite par la condition

(H.6) Il existe deux suites ϑ_n et β_n ainsi qu'une loi non dégénérée $\mathcal{D}$ telles que :
$\vartheta_n = o(\log(n/k_n)k_n^{1/2})$ et $\vartheta_n(\hat{\theta}_n - \beta_n - \theta) \xrightarrow{d} \mathcal{D}$.

Dans les deux situations, la suite (β_n) représente le biais asymptotique de l'estimateur de l'indice θ. Le théorème suivant établit la loi asymptotique des estimateurs de la famille $\{\hat{q}(\alpha_n, \hat{\theta}_n)\}$ dans les deux situations décrites ci-dessus. Ce résultat est énoncé et démontré dans [25, Théorème 2].

Théorème 8. *On se place sous le modèle (8) et on suppose que les conditions* **(H.1)** *et* **(H.2)** *sont satisfaites. Si* $\tau_n \to \tau \in]1, \infty[$ *alors :*

- *sous la condition* **(H.5)** *et si* $k_n^{1/2} \log(n/k_n) b(\log(n/k_n)) \to 0$,

$$\log(n/k_n) k_n^{1/2} \tau^{-\beta_n} \left(\frac{\hat{q}(\alpha_n, \hat{\theta}_n)}{q(\alpha_n)} - \tau_n^{\beta_n} \right) \xrightarrow{d} \mathcal{N}(0, \theta^2).$$

- *sous la condition* **(H.6)** *et si* $\vartheta_n b(\log(n/k_n)) \to 0$,

$$\frac{\vartheta_n}{\log(\tau)} \tau^{-\beta_n} \left(\frac{\hat{q}(\alpha_n, \hat{\theta}_n)}{q(\alpha_n)} - \tau_n^{\beta_n} \right) \xrightarrow{d} \mathcal{D}.$$

Le meilleur estimateur du quantile extrême $q(\alpha_n)$ est obtenu en utilisant un estimateur de θ satisfaisant l'hypothèse **(H.5)** avec $\beta_n = 0$. Malheureusement, à notre connaissance, un tel estimateur de θ n'existe pas. A titre d'exemple, l'estimateur de θ proposé par Broniatowski [12] satisfait la condition **(H.5)** mais avec un biais asymptotique β_n non nul. Pour la grande majorité des estimateurs de l'indice θ (notamment ceux introduits dans le paragraphe 3.1), c'est la condition **(H.6)** qui est satisfaite avec

un biais asymptotique pouvant être annulé.
La condition $\tau_n \to \tau \in]1, \infty[$ implique que l'on peut choisir un ordre α_n proportionnel à $n^{-\tau}$. Ainsi, plus τ est grand plus le quantile estimable est extrême. En contre partie, une grande valeur de τ augmente la variance asymptotique de l'estimateur dans les deux situations.
On peut montrer que lorsque la condition **(H.6)** est satisfaite avec un biais asymptotique $\beta_n = 0$, la vitesse de convergence de $\hat{q}(\alpha_n, \hat{\theta}_n)$ est de l'ordre de $(\log(n))^{-\rho-\epsilon}$ où $\epsilon \in]0, -\rho[$ peut être choisi aussi petit que l'on veut.

3.2.2 Un estimateur de $q(\alpha_n)$ débiaisé

Pour proposer un estimateur débiaisé du quantile extrême $q(\alpha_n)$, on se base sur le résultat suivant établi dans [17] : sous la condition **(H.2)**, si $\tau_n \to \tau \in]1, \infty[$, on a lorsque $n \to \infty$

$$q(\alpha_n) \sim q(k_n/n)\tau_n^{\theta} \exp\{b(\log(n/k_n))K_\rho(\tau_n)\}.$$

En estimant $q(k_n/n)$ par la statistique d'ordre $X_{n-k_n+1,n}$, θ par $\hat{\theta}_n^D$ (voir équation (18)), $b(\log(n/k_n))$ par $\hat{b}(\log(n/k_n))$ (voir équation (19)) et ρ par un estimateur $\hat{\rho}_n$, on obtient l'estimateur

$$X_{n-k_n+1,n}\tau_n^{\hat{\theta}_n^D} \exp\{\hat{b}(\log(n/k_n))K_{\hat{\rho}_n}(\tau_n)\}.$$

Si on néglige le terme de correction $\exp\{\hat{b}(\log(n/k_n))K_{\hat{\rho}_n}(\tau_n)\}$ dans l'expression ci-dessus, on retrouve l'estimateur non débiaisé $\hat{q}(\alpha_n, \hat{\theta}_n^D)$ appartenant à la famille $\mathcal{Q}_{\alpha_n}$ (voir équation (20)). Concernant le paramètre ρ, plusieurs estimateurs ont été proposés pour des modèles différents (citons les travaux de Gomes [35], Gomes *et al.* [36], Feuerverger *et al.* [24] Peng *et al.* [44] et Beirlant *et al.* [5]). Le résultat suivant (voir [17, Théorème 1]), montre que l'on peut remplacer $\hat{\rho}_n$ par une valeur arbitraire $\rho^\sharp < 0$ et obtenir un estimateur

$$\hat{q}^D(\alpha_n) = X_{n-k_n+1,n}\tau_n^{\hat{\theta}_n^D} \exp\{\hat{b}(\log(n/k_n))K_{\rho^\sharp}(\tau_n)\}$$

asymtotiquement normal.

Théorème 9. *On se place sous le modèle (8) et on suppose que les conditions* **(H.1)** *et* **(H.2)** *sont satisfaites. Supposons que la fonction $b(.)$ est telle que $x|b(x)| \to \infty$ lorsque $x \to \infty$ avec*

$$\frac{k_n^{1/2}}{\log(n/k_n)} b(\log(n/k_n)) \to \tilde{\Lambda} \in \mathbb{R}.$$

Supposons de plus que, si $\tilde{\Lambda} = 0$, $\dfrac{\log^2(k_n)}{\log(n/k_n)} \to 0$ et $\dfrac{k_n^{1/2}}{\log(n/k_n)} \to \infty$. Alors, si $\tau_n \to \tau \in]1, \infty[$ on a :

$$\frac{k_n^{1/2}}{\log(n/k_n)} \left(\frac{\hat{q}^D(\alpha_n)}{q(\alpha_n)} - 1 \right) \xrightarrow{d} \mathcal{N}(\tilde{\Lambda}\mu(\tau), \theta^2\sigma^2(\tau)),$$

avec $\sigma^2(\tau) = (K_{\rho^\sharp}(\tau) - \log(\tau))^2$ et $\mu(\tau) = (K_{\rho^\sharp}(\tau) - K_\rho(\tau))^2$.

Si $\tilde{\Lambda} \neq 0$ et si $\rho^\sharp = \rho$ alors l'estimateur $\hat{q}^D(\alpha_n)$ est sans biais avec une vitesse de convergence de l'ordre de $\log^{-\rho^\sharp}(n)\ell^*(\log(n))$ où $\ell^*(.)$ est une fonction à variations lentes. Cette vitesse est meilleure que celle obtenue pour les estimateurs de la famille $\mathcal{Q}_{\alpha_n}$ lorsque $\hat{\theta}_n$ satisfait l'hypothèse **(H.6)** avec un biais asymptotique $\beta_n = 0$. Evidemment un mauvais choix de $\rho^\sharp$ conduit à un estimateur du quantile extrême biaisé. Notons cependant que les lois à queue de type Weibull usuelles (loi normale, Gamma) ont un paramètre du second ordre $\rho = -1$. En pratique, on prendra donc une valeur $\rho^\sharp$ égale à -1.

4 Pour aller plus loin

Dans cet article, nous nous sommes intéressés à la famille des lois à queue de type Weibull. Ce type de lois interviennent dans de nombreuses applications (hydrologie notamment), des publications récentes

leur sont consacrées. Citons par exemple Dierckx *et al.* [18] qui utilisent la moyenne des excès au-dessus d'un seuil pour estimer l'indice de queue de Weibull et Goegebeur *et al.* [34] qui proposent des tests d'adéquation pour ces types de lois.

Nous avons étudié plusieurs estimateurs de l'indice de queue et de quantiles extrêmes. On aura pu remarquer la grande similarité entre les estimateurs de l'indice de queue de Weibull (équation (10) par exemple) avec l'estimateur de Hill (3) dédié à l'indice des valeurs extrêmes dans le cas de lois à queues lourdes. Pour expliquer ce phénomène, une famille de lois englobant notamment les lois du domaine d'attraction de Fréchet et les lois à queue de type Weibull a été introduite récemment [28]. La fonction de survie de ces lois est donnée par

$$\bar{F}(x) = \exp\big(-K_\tau^{-1}(\log(x^{1/\theta}\ell(x))\big), \tag{21}$$

où $\theta > 0$, $\ell(.)$ est une fonction à variations lentes et on rappelle que

$$K_\tau(x) = \int_1^x u^{\tau-1}du,\ \tau \in [0,1].$$

Ainsi, si $\tau = 0$, $\bar{F}(.)$ est la fonction de survie d'une loi à queue de type Weibull d'indice θ. Si $\tau = 1$, la fonction de survie est celle d'une loi du domaine d'attraction de Fréchet avec pour indice des valeurs extrêmes θ. Les valeurs intermédiaires de τ permettent d'inclure par exemple la loi lognormale.

Si τ est connu, le paramètre θ du modèle (21) est estimé par

$$\check{\theta}_n = \frac{1}{\mu_{1,\tau}(\log(n/k_n))}\frac{1}{k_n-1}\sum_{i=1}^{k_n-1}(\log(X_{n-i+1,n}) - \log(X_{n-k_n+1,n})),$$

avec pour $t > 0$,

$$\mu_{1,\tau}(t) = \int_0^\infty (K_\tau(x+t) - K_\tau(t))\,\mathrm{e}^{-x}dx.$$

La normalité asymptotique de cet estimateur est établie dans [28, Théorème 1]. On retrouve comme cas particuliers les résultats de normalité de l'estimateur de l'indice des valeurs extrêmes proposé par Hill [38] et l'estimateur de l'indice de queue de Weibull proposé par Beirlant *et al.* [8]. Un estimateur des quantiles extrêmes est également introduit :

$$\check{q}_n(\alpha_n) = X_{n-k_n+1,n}\exp\left(\check{\theta}_n\left[K_\tau(\log(1/\alpha_n)) - K_\tau(\log(n/k_n))\right]\right).$$

et sa normalité asymptotique est établie dans [28, Théorème 2]. En pratique, le paramètre τ est inconnu. Un estimateur est proposé dans [43]. Ses propriétés asymptotiques sont établies ainsi que les conséquences de son estimation sur la loi limite des estimateurs $\check{\theta}_n$ et $\check{q}_n(\alpha_n)$. Ces travaux ouvrent la porte à la construction de tests d'hypothèses pour les queues de distributions. Pour un jeu de données réelles, on pourrait notamment décider s'il est issu d'une loi à queue lourde ou d'une loi à queue de type Weibull.

References

[1] M. Abramowitz and J.A. Stegan. *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables*. Dover, New York, 1972.

[2] V. Asimit, D. Li, and L. Peng. Pitfalls in using Weibull tailed distributions. *Journal of Statistical Planning and Inference*, 140:2018–2024, 2010.

[3] J. Beirlant, C. Bouquiaux, and B. Werker. Semiparametric lower bounds for tail index estimation. *Journal of Statistical Planning and Inference*, 136:705–729, 2006.

[4] J. Beirlant, M. Broniatowski, J.L. Teugels, and P. Vynckier. The mean residual life function at great age: applications to tail estimation. *Journal of Statistical Planning and Inference*, 45:21–48, 1995.

[5] J. Beirlant, G. Dierckx, Y. Goegebeur, and G. Matthys. Tail index estimation and an exponential regression model. *Extremes*, 2:177–200, 1999.

[6] J. Beirlant, G. Dierckx, A. Guillou, and C. Stărică. On exponential representations of log-spacings of extreme order statistics. *Extremes*, 5:157–180, 2002.

[7] J. Beirlant and J.L. Teugels. Modelling large claims in non-life insurance. *Insurance: Mathematics and Economics*, 11:17–29, 1992.

[8] J. Beirlant, J.L. Teugels, and P. Vynckier. *Practical analysis of extreme values*. Leuven University Press, Leuven, Belgium, 1996.

[9] M. Berred. Record values and the estimation of the Weibull tail-coefficient. *Comptes-Rendus de l'Académie des Sciences*, T. 312, Série I:943–946, 1991.

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

[11] E. Brodin and H. Rootzén. Univariate and bivariate gpd methods for predicting extreme wind storm losses. *Insurance: Mathematics and Economics*, 44:345–356, 2009.

[12] M. Broniatowski. On the estimation of the Weibull tail coefficient. *Journal of Statistical Planning and Inference*, 35:349–366, 1993.

[13] S. Csörgő, P. Deheuvels, and D.M. Mason. Kernel estimates of the tail index of a distribution. *The Annals of Statistics*, 13:1050–1077, 1985.

[14] S. Csörgő and L. Viharos. Estimating the tail index. In B. Szyszkowicz, editor, *Asymptotic Methods in Probability and Statistics*, pages 833–881. TEST, North-Holland, Amsterdam, 1998.

[15] L. de Haan and A. Ferreira. *Extreme value theory: an introduction*. Springer, 2006.

[16] J. Diebolt, L. Gardes, S. Girard, and A. Guillou. Bias-reduced estimators of the Weibull tail-coefficient. *Test*, 17:311–331, 2008.

[17] J. Diebolt, L. Gardes, S. Girard, and A. Guillou. Bias-reduced extreme quantiles estimators of Weibull tail-distributions. *Journal of Statistical Planning and Inference*, 138:1389–1401, 2008.

[18] G. Dierckx, J. Beirlant, D. De Waal, and A. Guillou. A new estimation method for Weibull-type tails based on the mean excess function. *Journal of Statistical Planning and Inference*, 139(6):1905–1920, 2009.

[19] O. Ditlevsen. Distribution arbitrariness in structural reliability. In Balkema, editor, *Structural Safety and Reliability*, pages 1241–1247. TEST, Rotterdam, 1998.

[20] H. Drees and E. Kaufmann. Selecting the optimal sample fraction in univariate extreme value estimation. *Stochastic process and Application*, 75:149–172, 1998.

[21] P. Embrechts, C. Klüppelberg, and T. Mikosch. *Modelling extremal events*. Springer, 1997.

[22] M. Falk. Some best parameter estimates for distributions with finite endpoint. *Statistics*, 27:115–125, 1995.

[23] M. Falk and F. Marohn. Efficient estimation of the shape parameter in Pareto models with partially known scale. *Statistics & Decisions*, 15:229–239, 1997.

[24] A. Feuerverger and P. Hall. Estimating a tail exponent by modelling departure from a Pareto distribution. *The Annals of Statistics*, 27:760–781, 1999.

[25] L. Gardes and S. Girard. Estimating extreme quantiles of Weibull tail-distributions. *Communication in Statistics - Theory and Methods*, 34:1065–1080, 2005.

[26] L. Gardes and S. Girard. Comparison of Weibull tail-coefficient estimators. *REVSTAT - Statistical Journal*, 4(2):163–188, 2006.

[27] L. Gardes and S. Girard. Estimation of the Weibull tail-coefficient with linear combination of upper order statistics. *Journal of Statistical Planning and Inference*, 138:1416–1427, 2008.

[28] L. Gardes, S. Girard, and A. Guillou. Weibull tail-distributions revisited: a new look at some tail estimators. *Journal of Statistical Planning and Inference*, 141(1):429–444, 2011.

[29] J.L. Geluk and L. De Haan. *Regular variation, extensions and Tauberian theorems*. Center for Mathematics and Computer Science, Amsterdam, Netherlands, 1987.

[30] S. Girard. A Hill type estimate of the Weibull tail-coefficient. *Communication in Statistics - Theory and Methods*, 33(2):205–234, 2004.

[31] S. Girard, A. Guillou, and G. Stupfler. Estimating an endpoint with high order moments in the Weibull domain of attraction. *Statistics and Probability Letters*, 82:2136–2144, 2012.

[32] B. Gnedenko. Sur la distribution limite du terme maximum d'une série aléatoire. *The Annals of Mathematics*, 44:423–453, 1943.

[33] Y. Goegebeur, J. Beirlant, and T. de Wet. Generalized kernel estimators for the Weibull-tail coefficient. *Communications in Statistics - Theory and Methods*, 39(20):3695–3716, 2010.

[34] Y. Goegebeur and A. Guillou. Goodness-of-fit testing for Weibull-type behavior. *Journal of Statistical Planning and Inference*, 140(6):1417–1436, 2010.

[35] M.I. Gomes. Asymptotic unbiased estimators of the tail index based on external estimation of the second order parameter. *Extremes*, 5(1):5–31, 2002.

[36] M.I. Gomes, M.J. Martins, and M. Neves. Improving second order reduced bias extreme value index estimation. *REVSTAT - Statistical Journal*, 5(2):177–207, 2007.

[37] P. Hall and B.U. Park. New methods for bias correction at endpoints and boundaries. *The Annals of Statistics*, 30(5):1460–1479, 2002.

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

[39] M. Kratz and S. Resnick. The qq-estimator and heavy tails. *Stochastic Models*, 12:699–724, 1996.

[40] O.V. Lepski, E. Mammen, and V.G. Spokoiny. Optimal spatial adaption to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *The Annals of Statistics*, 25:929–947, 1997.

[41] D.M. Mason. Asymptotic normality of linear combinations of order statistics with a smooth score function. *The Annals of Statistics*, 9(4):899–908, 1981.

[42] C. Mercadier and P. Soulier. Optimal rates of convergence in the Weibull model based on kernel-type estimators. *Statistics and Probability Letters*, 82:548–556, 2011.

[43] J. El Methni, L. Gardes, S. Girard, and A. Guillou. Estimation of extreme quantiles from heavy and light tailed distributions. *Journal of Statistical Planning and Inference*, 142(10):2735–2747, 2012.

[44] L. Peng and Q. Yongcheng. Estimating the first and second order parameters of a heavy tailed distribution. *Australian and New Zealand Journal of Statistics*, 46(2):305–312, 2004.

[45] S.I. Resnick. *Extreme values, regular variation and point processes*. Springer Series in Operations Research and Financial Engineering, 1987.

[46] H. Rootzén and T. Tajvidi. Can losses caused by wind storms be predicted from meteorological observations? *Scandinavian Actuarial Journal*, 5:162–175, 2001.

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

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