REVSTAT – Statistical Journal
Volume 0, Number 0, Month 0000, 000-000
https://doi.org/00.00000/revstat.v00i0.000

# A refined extreme quantile estimator for Weibull tail-distributions

Authors: Jonathan El Methni ✉
– Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble France
jonathan.el-methni@univ-grenoble-alpes.fr

Stéphane Girard
– Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble France
stephane.girard@inria.fr

Received: Month 0000 Revised: Month 0000 Accepted: Month 0000

Abstract:

- We address the estimation of extreme quantiles of Weibull tail-distributions. Since such quantiles are asymptotically larger than the sample maximum, their estimation requires extrapolation methods. In the case of Weibull tail-distributions, classical extreme-value estimators are numerically outperformed by estimators dedicated to this set of light-tailed distributions. The latter estimators are based on two key quantities: an order statistic to estimate an intermediate quantile and an estimator of the Weibull tail-coefficient used to extrapolate. The common practice is to select the same intermediate sequence for both estimators. We show how an adapted choice of two different intermediate sequences leads to a reduction of the asymptotic bias associated with the resulting refined estimator. This analysis is supported by an asymptotic normality result associated with the refined estimator. A data-driven method is introduced for the practical selection of the intermediate sequences and our approach is compared to three estimators of extreme quantiles on simulated data. An illustration on a real data set of daily wind measures is also provided.

Keywords:

- *Extreme quantile, bias reduction, Weibull tail-distribution, extreme-value statistics, asymptotic normality.*

AMS Subject Classification:

- 60G70, 62G32, 62G20.

✉ Corresponding author

## 1. INTRODUCTION

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables with cumulative distribution function $F$ and let $X_{1,n} \leq \ldots \leq X_{n,n}$ denote the associated order statistics. We consider the case where $F$ belongs to the family of Weibull tail-distributions [12]:

**(A.1)** $F$ is twice differentiable and $F(\cdot) = 1 - \exp(-H(\cdot))$ with $V(t) := H^{\leftarrow}(t) = t^{\theta}\ell(t)$ for all $t > 0$, where $\theta > 0$ is called the Weibull tail-coefficient and where $\ell$ is a (positive) slowly-varying function *i.e.* $\ell(tx)/\ell(x) \to 1$ as $x \to \infty$ for all $t > 0$.

Here, and in the following, $\Phi^{\leftarrow}(\cdot) = \inf\{x \in \mathbb{R}, \Phi(x) > \cdot\}$ denotes the generalized inverse of an increasing function $\Phi$. The inverse cumulative hazard function $V$ is said to be regularly-varying at infinity with index $\theta$ and this property is denoted by $V \in \mathcal{RV}_{\theta}$, see [11] for a detailed account on this topic. The shape parameter $\theta$ is referred to as the Weibull tail-coefficient. Weibull tail-distributions are part of the Gumbel maximum domain of attraction, *i.e.* with extreme-value index $\gamma = 0$, see [19, Proposition 2(ii)], and as such, are light-tailed distributions. They include for instance exponential ($\theta = 1$), Gamma ($\theta = 1$), logistic ($\theta = 1$), Normal ($\theta = 1/2$) and Weibull distributions ($\theta$ is the inverse of the shape parameter), see [20, Table 1]. We refer to [9] for an application to the modeling of large claims in non-life insurance and to [26] for an analysis of neural networks distributional properties.

Dedicated methods have been proposed to estimate the Weibull tail-coefficient $\theta$ since the relevant information is localised in the extreme upper part of the sample. Most approaches rely on the $k_n$ upper order statistics $X_{n-k_n+1,n}, \ldots, X_{n,n}$ where $k_n \to \infty$ as $n \to \infty$. Note that, since $\theta$ is defined through a tail behavior, the associated estimator should only use the extreme-values of the sample and thus the extra condition $k_n/n \to 0$ is required. More specifically, recent estimators are based on the log-spacings between the $k_n$ upper order statistics [10, 17, 19, 20] or on the mean excess function [6, 7, 8]. The introduction of kernel based weights has been investigated for both approaches, see [18] for the log-spacings case and [21] for the mean excess function framework. A bias reduction method adapted to the estimation of the Weibull tail-coefficient is proposed in [14] and the adaptation to random censoring is achieved in [28].

We address the problem of estimating extreme quantiles of Weibull tail-distributions. Recall that an extreme quantile $q(\alpha_n)$ of order $\alpha_n$ is defined by $q(\alpha_n) = F^{\leftarrow}(1 - \alpha_n)$ with $n\alpha_n \to 0$ as $n \to \infty$. The latter condition implies that $q(\alpha_n)$ is almost surely asymptotically larger than $X_{n,n}$, the sample maximum. It is shown in [16] that classical extreme-value estimators of such large quantiles are numerically outperformed by estimators dedicated to Weibull tail-distributions [15], see also Lemma 1.1 in the Appendix for a theoretical argument.

The latter methods estimate $q(\alpha_n)$ by combining two ingredients: an order statistic $X_{n-k_n+1,n}$ and an estimator of the Weibull tail-coefficient $\theta$ used to extrapolate from this anchor point.

In this work, we show that the biases associated with the previous extrapolation method and the estimator of $\theta$ may asymptotically cancel out in the extreme quantile estimator thanks to an appropriate tuning of the number of upper order statistics involved in the Weibull tail-coefficient estimator. The construction of the resulting estimator is presented in Section 2 and an asymptotic normality result is provided, emphasizing that the proposed extreme quantile estimator is asymptotically less biased than the original one [16]. Its performances are illustrated on simulated data in Section 3 and compared to three state-of-the-art competitors [8, 15, 16]. An illustration on a real data set of daily wind measures is provided in Section 4. Finally, a small conclusion is proposed in Section 5 and the proofs are postponed to the Appendix.

## 2. A REFINED ESTIMATOR OF THE EXTREME QUANTILE

### 2.1. Extreme quantile estimators

Weibull-tail estimators of the extreme quantile $q(\alpha_n)$ rely on an intermediate quantile $q(k_n/n)$ where $(k_n)$ is an intermediate sequence of integers *i.e.* such that $k_n \in \{1,\dots,n-1\}$, $k_n \to \infty$ and $k_n/n \to 0$ as $n \to \infty$, see for instance [15, 16]. Indeed, in view of (**A.1**), one has

$$\frac{q(\alpha_n)}{q(k_n/n)} = \frac{V(\log(1/\alpha_n))}{V(\log(n/k_n))} \simeq \left(\frac{\log(1/\alpha_n)}{\log(n/k_n)}\right)^{\theta} =: \tau_n^{\theta}, \tag{2.1}$$

as $n \to \infty$, where $\tau_n = \log(1/\alpha_n)/\log(n/k_n)$ is the (logarithmic) extrapolation factor. From an intuitive point of view, an extreme quantile can thus be approximated by multiplying an intermediate quantile by an appropriate extrapolation term: $q(\alpha_n) \simeq q(k_n/n)\tau_n^{\theta}$. The intermediate quantile $q(k_n/n)$ can then be estimated by its empirical counterpart $X_{n-k_n+1,n}$ while the extrapolation term depends on the tail heaviness through $\theta$ which has to be estimated as well. Following the ideas of [4], we propose a refined Weissman [27] type estimator:

$$\hat{q}_n(\alpha_n, k_n, k_n') = X_{n-k_n+1,n}\left(\frac{\log(1/\alpha_n)}{\log(n/k_n)}\right)^{\hat{\theta}_n(k_n')} = X_{n-k_n+1,n}\,\tau_n^{\hat{\theta}_n(k_n')}, \tag{2.2}$$

with $\hat{\theta}_n(k_n')$ an estimator of $\theta$ depending on another intermediate sequence $(k_n')$. Let us focus on the estimator introduced in [17]:

$$\hat{\theta}_n^{\mathrm{RSH}}(k_n') = \frac{1}{\mu(\log(n/k_n'))}\frac{1}{k_n'}\sum_{i=1}^{k_n'}(\log X_{n-i+1,n} - \log X_{n-k_n'+1,n}), \tag{2.3}$$

with, for $t > 0$, $\mu(t) = \int_0^{+\infty} \log\left(1 + \frac{x}{t}\right) e^{-x} dx = e^t E_1(t)$, where $E_1$ is the Exponential integral function [2, Page 228]. This estimator is motivated by the remark that, in view of (2.1), the log-spacings between two quantiles are approximately proportional to $\theta$. This property is also used in the real data application (see the top-right panel of Figure 4) to visually check the Weibull-tail assumption. Clearly, $\hat{\theta}_n^{\mathrm{RSH}}(\cdot)$ can be interpreted as a rescaled Hill estimator since

$$\hat{\theta}_n^{\mathrm{RSH}}(k_n') = \frac{\hat{\gamma}_n^{\mathrm{H}}(k_n')}{\mu(\log(n/k_n'))}, \tag{2.4}$$

where $\hat{\gamma}_n^{\mathrm{H}}(\cdot)$ is the well-known Hill estimator [25] of the extreme-value index $\gamma$.

Let us note, when $k_n' = k_n$, one recovers the extreme quantile estimator for Weibull tail-distributions introduced in [16]. In the next paragraph, the asymptotic normality of $\hat{q}_n(\alpha_n, k_n, k_n')$ is established, and it is shown that choosing $k_n' \neq k_n$ can yield better results from an asymptotic point of view. A similar phenomenon occurs in the estimation of the endpoint of a distribution in the Weibull maximum domain of attraction, see [1] for details. We also refer to [4] for the estimation of the tail-index in the Fréchet maximum domain of attraction.

## 2.2. Asymptotic analysis

The study of the limit distribution of $\hat{q}_n(\alpha_n, k_n, k_n')$ requires a second-order condition on the slowly-varying function $\ell$ introduced in (**A**.**1**):

**(A.2)** There exist $\rho \leq 0$ and $b(t) \to 0$ as $t \to \infty$, with ultimately constant sign, such that uniformly locally on $x \geq 1$,

$$\lim_{t\to\infty} \frac{1}{b(t)} \log\left(\frac{\ell(tx)}{\ell(t)}\right) = K_\rho(x) := \int_1^x u^{\rho-1} du.$$

It can be shown that necessarily $|b| \in \mathcal{RV}_\rho$. The second-order Weibull parameter $\rho \leq 0$ tunes the rate of convergence of the ratio $\ell(tx)/\ell(t)$ to 1. The closer $\rho$ is to 0, the slower is the convergence. Condition **(A.2)** is the cornerstone in all proofs of asymptotic normality for extreme-value estimators. Again, we refer to [20, Table 1] for $\rho$ parameters associated with usual Weibull tail-distributions. Our first result is a refinement of [17, Corollary 3.1]. It provides an asymptotic normality result for the extreme quantile estimator (2.2) based on two intermediate sequences $(k_n)$ and $(k_n')$.

**Theorem 2.1.** *Assume* **(A.1)** *and* **(A.2)** *hold. Let* $(k_n)$ *and* $(k_n')$ *be two intermediate sequences and introduce* $(\alpha_n)$ *a probability sequence such that* $\alpha_n \to 0$ *as* $n \to \infty$. *Suppose, as* $n \to \infty$,

**(i)** $\sqrt{k_n'}\, b(\log(n/k_n')) \to \lambda \in \mathbb{R}$,

**(ii)** $\log(n/k_n')/\log(n/k_n) \to \beta \geq 1$,

**(iii)** $\tau_n \to \tau > \beta$.

*Then, as* $n \to \infty$,

$$\sqrt{k_n'}\left(\frac{\hat{q}_n(\alpha_n, k_n, k_n')}{q(\alpha_n)} - 1\right) \xrightarrow{d} \mathcal{N}\left(\lambda(\log(\tau) - \beta^{-\rho}K_\rho(\tau)), (\theta\log\tau)^2\right). \tag{2.5}$$

Let us first remark that condition (i) implies $\log(n/k_n') \sim \log(n)$ as $n \to \infty$ (see [17, Lemma 5.1]), then condition (ii) yields $\log(n/k_n) \sim \log(n)/\beta$ and therefore condition (iii) can be rewritten as $\log(1/\alpha_n) \sim (\tau/\beta)\log(n)$ as $n \to \infty$. As a consequence, the condition $\tau > \beta$ in (iii) implies $n\alpha_n \to 0$ as $n \to \infty$ which, in turns, implies that $q(\alpha_n)$ is an extreme quantile.

It follows from (2.5) that the asymptotic bias associated with $\hat{q}_n(\alpha_n, k_n, k_n')$ is given by

$$\begin{aligned}\left(\log\tau - \beta^{-\rho}K_\rho(\tau)\right) b(\log(n/k_n')) &\sim (\beta^\rho\log(\tau) - K_\rho(\tau))\, b(\log(n/k_n)) \\ &=: B(\beta,\tau,\rho) b(\log(n/k_n)),\end{aligned}$$

since $|b| \in \mathcal{RV}_\rho$. It appears that each choice of $k_n'$ yields an associated constant $\beta$ in (ii) and thus a corresponding bias factor $B(\beta,\tau,\rho) = \beta^\rho\log(\tau) - K_\rho(\tau)$. From the theoretical point of view, two cases can be considered.

• The usual choice $k_n' = k_n$ yields $\beta = 1$ and one recovers [17, Corollary 3.1] as a particular case of Theorem 2.1. Moreover, for all $\tau > 1$, $\rho \leq 0$,

$$B(1,\tau,\rho) = \log(\tau) - K_\rho(\tau) \geq 0, \tag{2.6}$$

which is the (positive) bias factor associated with the extreme quantile estimator $\hat{q}_n(\alpha_n, k_n, k_n)$ investigated in [17]. Note that $\rho \mapsto B(1,\tau,\rho)$ is a decreasing function such that $B(1,\tau,0) = 0$ which is an unusual situation in extreme-value theory. For instance, the bias factor associated with the Weissman estimator [27] dedicated to heavy-tailed distributions is proportional to $1/(1-\rho)$ and increases with $\rho$, see [22, Theorem 3.2.5 and Theorem 4.3.8].

• The choice $\beta^\star(\tau,\rho) := (K_\rho(\tau)/\log(\tau))^{1/\rho}$ yields

$$B(\beta^\star(\tau,\rho),\tau,\rho) = 0. \tag{2.7}$$

The associated intermediate sequence is given by $k_n^\star(\tau,\rho) = \lfloor n(k_n/n)^{\beta^\star(\tau,\rho)}\rfloor$ and therefore the extreme quantile estimator $\hat{q}_n(\alpha_n, k_n, k_n^\star(\tau,\rho))$ is asymptotically unbiased. Note that this estimator cannot be used in practice since the second-order Weibull parameter $\rho$ is unknown.

Up to our knowledge, there is no estimator of the second-order Weibull parameter in the statistical literature. In practice, one can replace the true unknown value of $\rho$ by a misspecified value $y \leq 0$ in the above $\beta^\star(\tau,\rho)$ leading to

$$\beta^\star(\tau,y) = (K_y(\tau)/\log(\tau))^{1/y}, \tag{2.8}$$

$$k_n^\star(\tau,y) = \lfloor n(k_n/n)^{\beta^\star(\tau,y)} \rfloor, \tag{2.9}$$

$$\begin{aligned} B(\beta^\star(\tau,y),\tau,\rho) &= \beta^\star(\tau,y)^\rho \log(\tau) - K_\rho(\tau) \\ &= (K_y(\tau)/\log(\tau))^{\rho/y}\log(\tau) - K_\rho(\tau), \end{aligned} \tag{2.10}$$

with $\rho \leq 0$ and $\tau > 1$. This misspecification technique has been used both to deal with Pareto-type distributions ($\gamma > 0$), see for instance [13], and Weibull tail-distributions ($\gamma = 0$) [15]. Some properties of the intermediate sequence $k_n^\star(\tau,y)$ are given in the next Lemma.

**Lemma 2.1.** *Let $\beta^\star(\tau,y)$ and $k_n^\star(\tau,y)$ be defined by (2.8) and (2.9) respectively. Then, for all $\tau > 1$:*

**(i)** *$\beta^\star(\tau,y) \to 1$ as $y \to -\infty$ and $\beta^\star(\tau,\cdot)$ can be extended by continuity by setting $\beta^\star(\tau,0) := \sqrt{\tau}$.*

**(ii)** *$1 < \beta^\star(\tau,y) < \tau$ for all $y \leq 0$.*

**(iii)** *For all $y \leq 0$, $k_n^\star(\tau,y)$ is an increasing function of $k_n$, $k_n^\star(\tau,y) \leq k_n$ and $k_n^\star(\tau,y)/k_n \to 0$ as $n \to \infty$.*

**(iv)** *$k_n^\star(\tau,y)$ is a decreasing function of $y \in (-\infty,0)$.*

In particular, it appears in (iii) that the number of upper order statistics $k_n^\star(\tau,y)$ used in the Weibull tail-coefficient estimator should be asymptotically small compared to $k_n$ for all finite values of $y$. From (iv), this is all the more true as $y$ is large. When $y \to -\infty$, meaning that one does not take into account the bias, (i) shows that $k_n^\star(-\infty,\tau) = k_n$ is recovered as a limit case. These properties are illustrated on the left panel of Figure 1, where $k_n^\star$ is drawn as a function of $k_n$ for several values of $y$. The next Corollary shows that these choices indeed lead to a bias reduction in the estimation of the extreme quantile.

**Corollary 2.1.** *Assume* **(A.1)** *and* **(A.2)** *hold. Let $c > 0$, $\tau > 1$, $y \leq 0$, $\lambda \neq 0$ such that $\lambda b(\cdot)$ is ultimately positive, and $\beta^\star(\tau,y)$ be defined as in (2.8). Let $\alpha_n = c\, n^{-\tau/\beta^\star(\tau,y)}$, $k_n = \lfloor n\{\lambda^2/(nb^2(\log n))\}^{1/\beta^\star(\tau,y)} \rfloor$ and define $k_n^\star(\tau,y)$ as in (2.9).*

**(i)** *Then, as $n \to \infty$,*

$$\sqrt{k_n^\star(\tau,y)}\left(\frac{\hat{q}_n(\alpha_n,k_n,k_n^\star(\tau,y))}{q(\alpha_n)} - 1\right) \xrightarrow{d} \mathcal{N}\left(\lambda(\log(\tau) - (K_y(\tau)/\log(\tau))^{-\rho/y}K_\rho(\tau)), (\theta\log\tau)^2\right).$$

**(ii)** *Moreover, for all $\tau > 1$, $y \leq 0$:*

$$\begin{aligned} |B(\beta^\star(\tau, y), \tau, \rho)| &< B(1, \tau, \rho) \text{ for all } \rho < 0, \\ B(\beta^\star(\tau, y), \tau, 0) &= B(1, \tau, 0) = 0. \end{aligned}$$

Let us first highlight that $\sqrt{k_n^\star(\tau, y)} \sim \lambda/b(\log n)$ as $n \to \infty$ (see the proof of Corollary 2.1 in the Appendix) which is the (logarithmic) rate of convergence of usual extreme quantile estimators dedicated to Weibull tail-distributions, see for instance [15, Theorem 1]. In contrast, the rate of convergence of extreme quantile estimators is a power function of $n$ in the Fréchet maximum domain of attraction, see [22, Theorem 4.3.8 and Equation (3.2.10)] for Weissman estimator and [4, Corollary 2] for the associated refined version. This may be seen as a consequence of Lemma 1.1 in the Appendix where it is established that the second-order parameter associated with Weibull distributions is $\psi = 0$.

Surprisingly, as a consequence of Corollary 2.1(ii), the extreme quantile estimator $\hat{q}_n(\alpha_n, k_n, k_n^\star(\tau, y))$ computed with $k_n^\star(\tau, y)$ defined in (2.8) and (2.9) has a smaller asymptotic bias than the usual one $\hat{q}_n(\alpha_n, k_n, k_n)$ whatever the chosen value $y \leq 0$. Let us recall that, from (2.7), the theoretical best choice would be $y = \rho$. In practice, we use $y = \rho^\# = -1$ leading to $\beta^\star(\tau, -1) = \tau \log(\tau)/(\tau - 1)$. This "canonical" choice is also used in [15], see Section 3.2 hereafter. Let us stress that the use of a similar bias reduction method in the Fréchet maximum domain attraction [4] is not based on such a misspecification technique but requires the estimation of $\rho$.

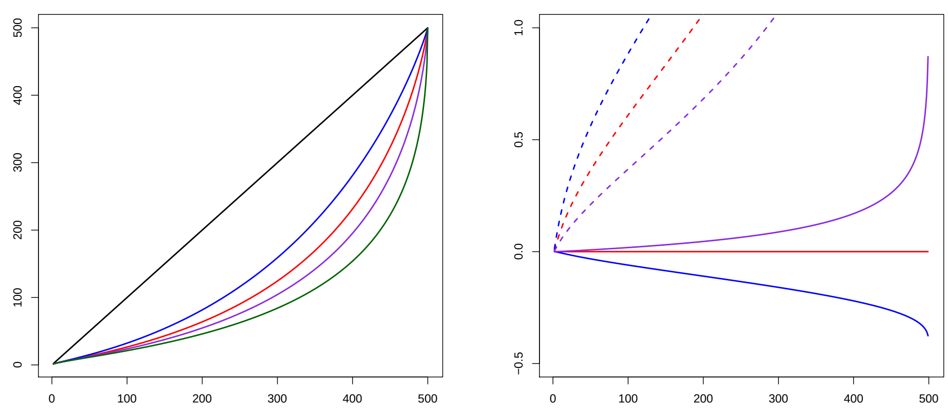


**Figure 1**: Left: Graphs of $k_n \in \{2, \ldots, 500\} \mapsto k_n^\star(\tau_n, y)$ for $y \in \{-\infty, -2, -1, -1/2, 0\}$ respectively in {black, blue, red, violet, green}. Right: graphs of $k_n \in \{2, \ldots, 500\} \mapsto B(1, \tau_n, \rho)$ (dotted lines) and $k_n \in \{2, \ldots, 500\} \mapsto B(\beta^\star(\tau_n, \rho^\# = -1), \tau_n, \rho)$ (solid lines) given in Equations (2.6) and (2.10), with $\rho \in \{-2, -1, -1/2\}$ respectively in {blue, red, violet}. On both panels: $\tau_n = \log(1/\alpha_n)/\log(n/k_n)$ with $\alpha_n = 1/n$ and $n = 500$.

Corollary 2.1(ii) is illustrated on the right panel of Figure 1 through the graphical comparison of the bias factors associated with $\beta = \beta^\star(\tau, -1)$ (refined Weibull-tail estimator) and $\beta = 1$ (usual Weibull-tail estimator [17]). It clearly appears that, from the theoretical point of view, the first choice yields smaller bias factors in absolute value than the second one. The performance of the refined estimator in practice is assessed on simulated data in the next Section.

## 3. VALIDATION ON SIMULATED DATA

The refined extreme quantile estimator is compared on simulated data to the original estimator [16] and to two other competitors described hereafter.

### 3.1. Experimental design

Let us consider the class of $\mathcal{D}(\zeta, \eta, a)$-distributions which is an adaptation of Hall's class [23, 24] to the Weibull-tail framework. In this family, the inverse cumulative hazard function is defined for all $x > 0$ by

$$V(x) := x^{1/\zeta}\left(1 + \frac{a}{\eta}x^{-\eta}\right),$$

with $a, \zeta, \eta > 0$ and $\zeta\eta \leq 1$. Under these conditions, the above class of distributions fulfills assumptions (**A.1**) and (**A.2**) with Weibull tail-coefficient $\theta = 1/\zeta$, second-order Weibull parameter $\rho = -\eta$, slowly-varying function $\ell(x) = 1 + (a/\eta)x^{-\eta}$ and $b(x) = -ax^{-\eta}$. Unlike classical distributions such as the (absolute) Normal distribution $\mathcal{N}(\mu, \sigma)$ ($\theta = 1/2$, $\rho = -1$ and $b(x) = \log(x)/(4x)$), the Gamma distribution $\mathcal{G}(\upsilon \neq 1, \lambda)$ ($\theta = 1$, $\rho = -1$ and $b(x) = (1-\upsilon)\log(x)/x$) and the Weibull distribution $\mathcal{W}(\upsilon, \lambda)$ ($\theta = 1/\upsilon$, $\rho = -\infty$ and $b(x) = 0$), it is thus possible to obtain $\mathcal{D}$-distributions with arbitrary Weibull tail-coefficient $\theta > 0$ and second-order Weibull parameter $\rho \in [-\theta, 0)$.

In the following, we set $\theta \in \{1/2, 3/4, \ldots, 5/2\}$, $\rho \in \{-5/2, -2, \ldots, -1/2\}$, $a = 10$ and focus on the only 25 situations of the $\mathcal{D}$-distribution where $\rho \geq -\theta$ to fulfill the constraint $\zeta\eta \leq 1$, see Table 1. We also consider 5 situations from the (absolute) Normal distribution $\mathcal{N}(\mu, \sigma = 1)$ with $\mu \in \{1, 3, 5, 7, 9\}$, 4 situations from the Gamma distribution $\mathcal{G}(\upsilon, \lambda = 1)$ with $\upsilon \in \{4, 6, 8, 10\}$ and 2 situations from the Weibull distribution $\mathcal{W}(\upsilon, \lambda)$ with $\upsilon = \lambda \in \{1/2, 2\}$. In each case, $N = 1,000$ replications of a data set of $n = 500$ i.i.d. realisations are simulated from the $25+5+4+2 = 36$ considered parametric models. Finally, the same two cases as in [15] are investigated for the order of the extreme quantile: $\alpha_n \in \{1/n^2, 1/n^4\}$. Summarizing, this experimental design includes $36 \times 2 = 72$ configurations.

| $(\rho,\theta)$ | 1/2 | 3/4 | 1 | 5/4 | 3/2 | 7/4 | 2 | 9/4 | 5/2 |
|---|---|---|---|---|---|---|---|---|---|
| $-\infty$ | $\mathcal{W}_1$ | | | | | | $\mathcal{W}_1$ | | |
| -5/2 | | | | | | | | | $\mathcal{D}_1$ |
| -2 | | | | | | | $\mathcal{D}_1$ | $\mathcal{D}_1$ | $\mathcal{D}_1$ |
| -3/2 | | | | | $\mathcal{D}_1$ | $\mathcal{D}_1$ | $\mathcal{D}_1$ | $\mathcal{D}_1$ | $\mathcal{D}_1$ |
| -1 | $\mathcal{N}_5$ | | $\mathcal{G}_4/\mathcal{D}_1$ | $\mathcal{D}_1$ | $\mathcal{D}_1$ | $\mathcal{D}_1$ | $\mathcal{D}_1$ | $\mathcal{D}_1$ | $\mathcal{D}_1$ |
| -1/2 | $\mathcal{D}_1$ | $\mathcal{D}_1$ | $\mathcal{D}_1$ | $\mathcal{D}_1$ | $\mathcal{D}_1$ | $\mathcal{D}_1$ | $\mathcal{D}_1$ | $\mathcal{D}_1$ | $\mathcal{D}_1$ |

**Table 1**: All considered configurations for $(\rho,\theta)$. The letter stands for the distribution and the subscript for the number of investigated situations. As an example, $\mathcal{N}_5$ corresponds to the (absolute) Normal distribution $\mathcal{N}(\mu,\sigma=1)$ where five cases are considered $\mu\in\{1,3,5,7,9\}$.

## 3.2. Competitors

The refined estimator dedicated to the estimation of extreme quantiles for Weibull tail-distributions is compared to three competitors. All three estimators share the same structure and rely on three quantities, *i.e.* the order statistic $X_{n-k_n+1,n}$, an extrapolation term and an estimator of the Weibull tail-coefficient.

Let us first consider the estimator (2.3) of the Weibull tail-coefficient introduced in [17]. The extreme quantile estimator proposed in [16] can be interpreted as a particular case of (2.2) with $k'_n=k_n$ and $\hat\theta_n(\cdot)=\hat\theta_n^{\mathrm{RSH}}(\cdot)$, see (2.4):

$$\hat q_n^{\mathrm{RSH}}(\alpha_n,k_n)=X_{n-k_n+1,n}\,\tau_n^{\hat\theta_n^{\mathrm{RSH}}(k_n)}. \tag{3.1}$$

More recently, an estimator of the Weibull tail-coefficient based on the mean excess function $t\mapsto m(t)=\mathbb{E}(X-t\mid X>t)$ has been introduced in [8]. In practice, the authors estimate $m(X_{n-j,n})$ for all $j\in\{1,\dots,k_n\}$ by its empirical counterpart:

$$\hat m_n(X_{n-j,n})=\frac{1}{j}\sum_{i=1}^{j}X_{n-i+1,n}-X_{n-j,n},$$

which leads to the following estimator of $\theta$ based on log-spacings between the mean excesses:

$$\hat\theta_n^{\mathrm{MEF}}(k_n)=\left(1-\frac{1}{\hat\gamma_n^{\mathrm{H}}(k_n)}\frac{1}{k_n}\sum_{j=1}^{k_n}\log\hat m_n(X_{n-j,n})-\log\hat m_n(X_{n-k_n-1,n})\right)^{-1}.$$

Letting $k'_n=k_n$ and $\hat\theta_n(\cdot)=\hat\theta_n^{\mathrm{MEF}}(\cdot)$ in (2.3) yields the following estimator of the extreme quantile:

$$\hat q_n^{\mathrm{MEF}}(\alpha_n,k_n)=X_{n-k_n+1,n}\,\tau_n^{\hat\theta_n^{\mathrm{MEF}}(k_n)}. \tag{3.2}$$

Up to our knowledge there exists only one bias-reduced extreme quantile estimator dedicated to Weibull tail-distributions. This estimator [15] is based on a least-squares approach and involves a bias-reduced estimator of the Weibull tail-coefficient proposed by the same authors [14]:

$$\hat{\theta}_n^{\mathrm{LSE}}(k_n) = \bar{Y}_{k_n} - \hat{b}(\log(n/k_n))\bar{x}_{k_n},$$

where

$$\bar{Y}_{k_n} = \frac{1}{k_n}\sum_{j=1}^{k_n} Y_j \ \text{ with } \ Y_j = j\log(n/j)(\log X_{n-j+1,n} - \log X_{n-j,n}),$$

$$\bar{x}_{k_n} = \frac{1}{k_n}\sum_{j=1}^{k_n} x_j \ \text{ with } \ x_j = \log(n/k_n)/\log(n/j),$$

and where

$$\hat{b}(\log(n/k_n)) = \sum_{j=1}^{k_n}(x_j - \bar{x}_{k_n})Y_j \Bigg/ \sum_{j=1}^{k_n}(x_j - \bar{x}_{k_n})^2 .$$

The associated extreme quantile estimator is defined as

$$\hat{q}_n^{\mathrm{LSE}}(\alpha_n, k_n) = X_{n-k_n+1,n}\,\tau_n^{\hat{\theta}_n^{\mathrm{LSE}}(k_n)}\exp\left(\widehat{b}(\log(n/k_n))K_{\widehat{\rho}_n}(\tau_n)\right). \tag{3.3}$$

The authors suggest to choose in practice $\widehat{\rho}_n = \rho^{\#} = -1$. This estimator features two bias corrections: a first one in the estimator of the Weibull tail-coefficient and a second one in the extrapolation term. This estimator is built under the assumption that $x|b(x)| \to \infty$ as $x \to \infty$ leading to the constraint $\rho \geq -1$. The latter assumption is fulfilled by the class of $\mathcal{D}(\zeta, \eta, a)$-distributions when $\eta \leq 1$.

Finally, recall that our estimator is given by

$$\hat{q}_n^{\mathrm{RWT}}(\alpha_n, k_n, k_n^\star(\tau_n, -1)) = X_{n-k_n+1,n}\,\tau_n^{\hat{\theta}_n^{\mathrm{RSH}}(k_n^\star(\tau_n,-1))}, \tag{3.4}$$

where $k_n^\star(\tau_n, -1) = \lfloor n(k_n/n)^{\beta^\star(\tau_n,-1)}\rfloor$ and $\beta^\star(\tau_n, -1) = \tau_n\log(\tau_n)/(\tau_n - 1)$. For the sake of simplicity, the above extreme quantile estimators (3.1)–(3.4) are respectively referred to as RSH, MEF, LSE and RWT in the sequel.

### 3.3. Selection of the intermediate sequence

All four considered extreme quantile estimators (RWT, RSH, LSE, MEF) depend on the intermediate sequence $k_n$. The selection of $k_n$ is a crucial point which has been widely discussed in the extreme-value literature. A new algorithm for the selection of $k_n$ is proposed in [4], basing on a bisection method inspired from random forests. The objective is to find the region with the smallest variance in a given series $\{Z_1, \ldots, Z_m\}$. The proposed method starts by randomly splitting the series into two parts, computes the variance in each sub-region and repeats the action in the one with smallest variance until getting a final single point, see [4, Algorithm 2]. The above procedure is embedded in a bootstrap technique, see [4, Algorithm 1], and the final $k_n^\dagger$ is selected by taking the median across the $T = 10,000$ bootstrap samples. In the simulations, $Z_j = \hat{q}(\alpha_n, k_{j,n})$, an estimator (RWT, RSH, LSE or MEF) of the extreme quantile at level $\alpha_n$ computed with the intermediate sequence $k_{j,n} \in \{15, 16, \dots, 3n/4\}$.

## 3.4. Results

The performance of the four extreme quantile estimators is assessed using the Mean absolute relative error:

$$\text{MARE}\left(\hat{q}_n(\alpha_n)\right) = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\hat{q}_n^{(i)}(\alpha_n, k_n^{(i,\dagger)})}{q_n(\alpha_n)} - 1\right|, \tag{3.5}$$

where $\hat{q}_n^{(i)}(\alpha_n, k_n^{(i,\dagger)})$ denotes the estimator computed on the $i$th replication, $i \in \{1, \dots, N = 1,000\}$ with the intermediate sequence $k_n^{(i,\dagger)}$ selected using the above described procedure. The computed MAREs are provided in Table 4 and Table 6 for $\alpha_n = 1/n^2$ and in Table 5 and Table 7 for $\alpha_n = 1/n^4$. The results are summarized in Table 2: We start by remarking that, as expected, the smaller the order $\alpha_n$ of the extreme quantile is, *i.e.* the more one extrapolates, the larger the error is. This is true for all four considered estimators. The proposed RWT estimator is the most accurate one in average since it provides the best results in 48% of cases. Let us remark that, since we fixed $\rho^\# = -1$, the RWT estimator performs best overall when $\rho$ is close to $-1$. The second most accurate estimator is LSE which provides the best results in 26% of the considered cases (19 out of 72 situations). As expected, and similarly to the RWT estimator, it performs well when $\rho = -1$. The RSH estimator shares similar performances with 25% of best results. It is remarkably efficient when $\rho = -\infty$ (all 4 situations) and more surprisingly when $\rho = -1/2$ where it obtains 14 best results. RSH performs well in this difficult case despite the fact that it does not benefit from a bias reduction. This unexpected performance may be explained by the relatively small bias factor, see the graph of $B(1, \cdot, -1/2)$ in the right panel of Figure 1. The four cases where RSH fails to obtain the best results when $\rho = -1/2$ correspond to a Weibull tail-coefficient $\theta$ smaller than 1. Finally, MEF yields very poor estimations (even in the strict Weibull case), with less than 2% of best results (only 1 situation). In particular, it does not give acceptable results (with MARE $\geq 1$) in 22% of the considered situations.

As an illustration, the median and MARE associated with the RWT, RSH and LSE estimators computed on a $\mathcal{D}(\zeta, \eta, a = 10)$-distribution for $\alpha_n = 1/n^2$ are depicted on Figure 2 and Figure 3 as functions of $k_n$. In Figure 2, the Weibull tail-coefficient is fixed to $\theta = 3/2$ and $\rho \in \{-1/2, -1, -3/2\}$ decreases (from top to bottom), while, in Figure 3, the second-order Weibull parameter is fixed to $\rho = -1$ and $\theta \in \{1, 3/2, 2\}$ increases (from top to bottom). In most of these situations, the RWT estimator has the smallest bias and thus the minimum value of the MARE is reached for larger values of $k_n$ than RSH and LSE. To conclude,

| $(\rho, \theta)$ | 1/2 | 3/4 | 1 | 5/4 | 3/2 | 7/4 | 2 | 9/4 | 5/2 |
|---|---|---|---|---|---|---|---|---|---|
| $-\infty$ | RSH | | | | | | RSH | | |
| -5/2 | | | | | | | | | LSE |
| -2 | | | | | | | RWT | RWT | RWT |
| -3/2 | | | | | **RWT** | RWT | RWT | RWT | RWT |
| -1 | RWT/LSE | | **RWT/LSE** | RWT | **RWT** | LSE/MEF | **LSE** | LSE | LSE |
| -1/2 | RWT | LSE | RSH | RSH | **RSH** | RSH | RSH | RSH | RSH |

**Table 2**: Summary of results obtained in Tables 4–7. Best estimator of the extreme quantiles $q(\alpha_n = 1/n^2)$ & $q(\alpha_n = 1/n^4)$ computed on simulated data from Weibull tail-distributions. The situations in bold are illustrated for the $\mathcal{D}(\zeta = 1/\theta, \eta = -\rho, a = 10)$-distribution on Figure 2 and Figure 3.

it appears on these experiments that, overall, the RWT estimator performs the best within the four considered estimators. One of its main competitors is LSE, which, similarly to RWT, considers the two sources of bias (associated with the Weibull tail-coefficient estimator and the extrapolation term).

## 4. ILLUSTRATION ON A REAL DATA SET

We study a data set of daily wind measures (in $m/s$) at Reims (France) from 01/01/1981 to 04/30/2011. For seasonality reasons, only the months from October to March are considered, resulting in $n = 5,371$ measures, see the top-left panel of Figure 4 for an histogram of the considered data. It is shown in [3] that the Weibull tail model represents fairly well the upper tail of these data. The goal is to estimate the extreme quantile $q(1/n)$ (with $1/n \simeq 1.86 \cdot 10^{-4}$) and to compare it to the maximum of the sample $x_{n,n} = 42.26\ m/s$.

To this end, the Weibull tail-coefficient is estimated first by $\hat{\theta}_n^{\text{RWT}}(k_n^\dagger) = \hat{\theta}_n^{\text{RSH}}(\hat{k}_n^\star) = 0.5597$, where $k_n^\dagger = 2,877$ has been selected following the procedure introduced in [4] and sketched in Subsection 3.3. This yields $\hat{k}_n^\star = \hat{k}_n^\star(\tau_n, \rho^\#) = 961$ where we set $\rho^\# = -1$. As a visual check, a Weibull quantile-quantile plot of the log-excesses $(\log X_{n-i+1,n} - \log X_{n-\hat{k}_n^\star+1,n})$ versus $(\log\log(n/i) - \log\log(n/\hat{k}_n^\star))$ for $i \in \{1, \ldots, \hat{k}_n^\star\}$ is drawn on the top-right panel of Figure 4. The relationship appearing in this plot is approximately linear, which constitutes an empirical evidence that the Weibull-tail assumption makes sense and that $\hat{k}_n^\star = 961$ is a reasonable choice for the estimation of the Weibull tail-coefficient. A line with the estimated value $\hat{\theta}_n^{\text{RWT}}(k_n^\dagger) = 0.5597$ as slope is added to the quantile-quantile plot highlighting the linear relationship. The function $k_n \mapsto \hat{\theta}_n^{\text{RSH}}(k_n)$ is plotted on the bottom-left panel of Figure 4, it features as a nice stability for all $k_n \in \{100, \ldots, 4000\}$. A similar procedure is carried out for the other three estimators to select $k_n^\dagger$. The two Weibull tail-coefficient estimators RSH and MEF that do not benefit from a bias reduction provide respectively the smallest and the largest estimation: $\hat{\theta}_n^{\text{RSH}}(k_n^\dagger) = 0.5017$ and $\hat{\theta}_n^{\text{MEF}}(k_n^\dagger) = 0.6693$, while the bias-reduced estimator LSE gives a value $\hat{\theta}_n^{\text{LSE}}(k_n^\dagger) = 0.6077$ close to the RWT estimate $\hat{\theta}_n^{\text{RWT}}(k_n^\dagger) = 0.5597$. These results are reported in Table 3 with the corresponding estimated extreme quantiles $\hat{q}_n(1/n)$. The estimates of the extreme quantile provided by RSH and MEF seem to respectively underestimate and overestimate $q(1/n)$ with respectively $\text{RSH}(1/n) = 33.89$ $m/s$ and $\text{MEF}(1/n) = 49.53$ $m/s$ while the sample maximum is $x_{n,n} = 42.26$ $m/s$. It appears that $\text{LSE}(1/n) = 37.08$ is significantly smaller than the sample maximum. Let us stress that the proposed refined estimator gives the closest estimate to the maximum value of the sample: $\text{RWT}(1/n) = 41.00$ $m/s$. Note that the behaviour of the RWT estimate is stable with respect to the choice of $\rho^\#$: $\text{RWT}(1/n) \in \{37.62, 41.00, 39.84\}$ when $\rho^\# \in \{-2, -1, -1/2\}$ even though $\rho^\# = -1$ seems to be the best option. Finally, both sample paths $k_n \mapsto \text{RWT}(1/n)$ and $k_n \mapsto \text{LSE}(1/n)$ enjoy a stable behaviour in a large neighbourhood of $k_n^\dagger$, see the bottom-right panel of Figure 4. As a conclusion, according to $\text{RWT}(1/n)$ estimate, one can expect a daily wind larger than 41.00 $m/s$ to occur in average once every 30 years during the October to March period.

| | RSH | RWT | LSE | MEF |
|---|---|---|---|---|
| $\hat{\theta}_n(k_n^\dagger)$ | 0.5017 | 0.5597 | 0.6077 | 0.6693 |
| $\hat{q}_n(1/n, k_n^\dagger)$ | 33.89 | 41.00 | 37.078 | 49.53 |
| $k_n^\dagger$ | 2,206 | 2,877 | 2,792 | 2,202 |

**Table 3**: Comparison of the four estimators on the daily wind data set: Estimates of the Weibull tail-coefficient $\theta$ and extreme quantile $q(1/n)$. The selected intermediate sequence $k_n^\dagger$ is also given for each estimator.

## 5. CONCLUSION

As a conclusion, the RWT estimator is an efficient tool for estimating extreme quantiles from Weibull tail-distributions. It relies on the ideas of [4], consisting in selecting carefully two intermediate sequences to reduce the asymptotic bias of a Weissman type estimator. In contrast to this previous work, the proposed approach does not rely on a preliminary estimate of the second-order parameter; Any negative value may be used, and does yield an asymptotic bias reduction, as shown in our theoretical results. Other surprising features of Weibull tail-distributions can be found in [5]. The proposed method provides satisfying results in our numerical experiments and outperforms all its competitors in half of the considered situations. This work could be extended by investigating the adaptation of this bias reduction principle to other estimators of extreme quantiles from Weibull tail-distributions.

| | RWT | RSH | LSE | MEF |
|---|---|---|---|---|
| $\theta = 1/2$ | | | | |
| $\rho = -1/2$ | **0.0133** | 0.0441 | 0.0203 | 0.0969 |
| $\theta = 3/4$ | | | | |
| $\rho = -1/2$ | 0.0702 | 0.0459 | **0.0412** | 0.2543 |
| $\theta = 1$ | | | | |
| $\rho = -1$ | **0.1125** | 0.3449 | 0.1814 | - |
| $\rho = -1/2$ | 0.1317 | **0.0640** | 0.0715 | 0.3015 |
| $\theta = 5/4$ | | | | |
| $\rho = -1$ | **0.1430** | 0.3386 | 0.1846 | 0.5964 |
| $\rho = -1/2$ | 0.1937 | **0.0857** | 0.1029 | 0.2283 |
| $\theta = 3/2$ | | | | |
| $\rho = -3/2$ | **0.2116** | 0.7095 | 0.3844 | - |
| $\rho = -1$ | **0.1874** | 0.3374 | 0.1900 | 0.4198 |
| $\rho = -1/2$ | 0.2517 | **0.1090** | 0.1332 | 0.2076 |
| $\theta = 7/4$ | | | | |
| $\rho = -3/2$ | **0.2470** | 0.6989 | 0.3831 | - |
| $\rho = -1$ | 0.2442 | 0.3330 | **0.1986** | 0.2705 |
| $\rho = -1/2$ | 0.3154 | **0.1349** | 0.1663 | 0.3809 |
| $\theta = 2$ | | | | |
| $\rho = -2$ | **0.3236** | 0.8833 | 0.4406 | - |
| $\rho = -3/2$ | **0.2869** | 0.6945 | 0.3833 | - |
| $\rho = -1$ | 0.2934 | 0.3311 | **0.2136** | 0.3168 |
| $\rho = -1/2$ | 0.3744 | **0.1586** | 0.1971 | 0.5915 |
| $\theta = 9/4$ | | | | |
| $\rho = -2$ | **0.3710** | 0.8783 | 0.4365 | - |
| $\rho = -3/2$ | **0.3274** | 0.6915 | 0.3847 | 0.6088 |
| $\rho = -1$ | 0.3401 | 0.3303 | **0.2301** | 0.4946 |
| $\rho = -1/2$ | 0.4359 | **0.1818** | 0.2262 | 0.7415 |
| $\theta = 5/2$ | | | | |
| $\rho = -5/2$ | 0.8958 | 0.9493 | **0.3713** | - |
| $\rho = -2$ | **0.4301** | 0.8754 | 0.4420 | - |
| $\rho = -3/2$ | **0.3824** | 0.6891 | 0.3874 | 0.4998 |
| $\rho = -1$ | 0.4050 | 0.3309 | **0.2526** | 0.7070 |
| $\rho = -1/2$ | 0.5086 | **0.2101** | 0.2629 | 0.8675 |

**Table 4**: MAREs associated with the four estimators of the extreme quantile $q(\alpha_n = 1/n^2)$ computed on simulated data from the $\mathcal{D}(\zeta = 1/\theta, \eta = -\rho, a = 10)$-distribution. The best result is emphasized in bold. MAREs larger than 1 are not reported.

| | RWT | RSH | LSE | MEF |
|---|---|---|---|---|
| $\theta = 1/2$ | | | | |
| $\rho = -1/2$ | **0.0363** | 0.0835 | 0.0498 | 0.0871 |
| $\theta = 3/4$ | | | | |
| $\rho = -1/2$ | 0.1049 | 0.0823 | **0.0685** | 0.2907 |
| $\theta = 1$ | | | | |
| $\rho = -1$ | **0.2797** | 0.5471 | 0.3746 | - |
| $\rho = -1/2$ | 0.1928 | **0.0964** | 0.1075 | 0.3572 |
| $\theta = 5/4$ | | | | |
| $\rho = -1$ | **0.2941** | 0.5408 | 0.1471 | 0.4404 |
| $\rho = -1/2$ | 0.2768 | **0.1201** | 0.1495 | 0.2738 |
| $\theta = 3/2$ | | | | |
| $\rho = -3/2$ | **0.3973** | 0.8691 | 0.6227 | - |
| $\rho = -1$ | **0.3316** | 0.5384 | 0.3743 | 0.3368 |
| $\rho = -1/2$ | 0.3548 | **0.1467** | 0.1923 | 0.2555 |
| $\theta = 7/4$ | | | | |
| $\rho = -3/2$ | **0.4242** | 0.8629 | 0.6135 | - |
| $\rho = -1$ | 0.3813 | 0.5359 | 0.3744 | **0.3286** |
| $\rho = -1/2$ | 0.4454 | **0.1785** | 0.2406 | 0.4611 |
| $\theta = 2$ | | | | |
| $\rho = -2$ | **0.4764** | 0.9609 | 0.6591 | - |
| $\rho = -3/2$ | **0.4616** | 0.8601 | 0.6109 | - |
| $\rho = -1$ | 0.4351 | 0.5340 | **0.3796** | 0.4767 |
| $\rho = -1/2$ | 0.5256 | **0.2088** | 0.2883 | 0.6809 |
| $\theta = 9/4$ | | | | |
| $\rho = -2$ | **0.5542** | 0.9591 | 0.6429 | - |
| $\rho = -3/2$ | **0.4977** | 0.8585 | 0.6088 | 0.6181 |
| $\rho = -1$ | 0.4869 | 0.5316 | **0.3878** | 0.6607 |
| $\rho = -1/2$ | 0.6005 | **0.2355** | 0.3304 | 0.8201 |
| $\theta = 5/2$ | | | | |
| $\rho = -5/2$ | - | 0.9865 | **0.5552** | - |
| $\rho = -2$ | **0.6308** | 0.9578 | 0.6388 | - |
| $\rho = -3/2$ | **0.5603** | 0.8568 | 0.6047 | 0.6947 |
| $\rho = -1$ | 0.5557 | 0.5298 | **0.4045** | 0.8317 |
| $\rho = -1/2$ | 0.7019 | **0.2698** | 0.3847 | 0.9216 |

**Table 5**: MAREs associated with the four estimators of the extreme quantile $q(\alpha_n = 1/n^4)$ computed on simulated data from the $\mathcal{D}(\zeta = 1/\theta, \eta = -\rho, a = 10)$-distribution. The best result is emphasized in bold. MAREs larger than 1 are not reported.

| | RWT | RSH | LSE | MEF |
|---|---|---|---|---|
| $\mathcal{N}(\mu, \sigma = 1)$ | | | | |
| $\mu = 1$ | 0.1396 | 0.2804 | **0.0798** | 0.1852 |
| $\mu = 3$ | 0.0776 | 0.0564 | **0.0471** | 0.2943 |
| $\mu = 5$ | 0.0553 | 0.0882 | **0.0440** | 0.2682 |
| $\mu = 7$ | **0.0334** | 0.0789 | 0.0366 | 0.1852 |
| $\mu = 9$ | **0.0361** | 0.0822 | 0.0381 | 0.1987 |
| $\mathcal{G}(\upsilon, \lambda = 1)$ | | | | |
| $\upsilon = 4$ | 0.1417 | 0.2240 | **0.1089** | 0.4930 |
| $\upsilon = 6$ | 0.1219 | 0.2270 | **0.1071** | 0.5040 |
| $\upsilon = 8$ | **0.0943** | 0.2095 | 0.1086 | 0.4581 |
| $\upsilon = 10$ | **0.1002** | 0.2158 | 0.1017 | 0.4758 |
| $\mathcal{W}(\upsilon, \lambda)$ | | | | |
| $\upsilon = \lambda = 1/2$ | 0.4932 | **0.2100** | 0.2593 | 0.8528 |
| $\upsilon = \lambda = 2$ | 0.1217 | **0.0510** | 0.0646 | 0.3324 |

**Table 6**: MAREs associated with the four estimators of the extreme quantile $q(\alpha_n = 1/n^2)$ computed on simulated data from classical Weibull tail-distributions. The best result is emphasized in bold.

| | RWT | RSH | LSE | MEF |
|---|---|---|---|---|
| $\mathcal{N}(\mu, \sigma = 1)$ | | | | |
| $\mu = 1$ | 0.1914 | 0.4430 | **0.1212** | 0.3009 |
| $\mu = 3$ | 0.1162 | 0.1001 | **0.0797** | 0.3354 |
| $\mu = 5$ | **0.0837** | 0.1584 | 0.0897 | 0.2507 |
| $\mu = 7$ | **0.0780** | 0.1565 | 0.0892 | 0.2270 |
| $\mu = 9$ | **0.0702** | 0.1477 | 0.0856 | 0.1895 |
| $\mathcal{G}(\upsilon, \lambda = 1)$ | | | | |
| $\upsilon = 4$ | 0.2206 | 0.3616 | **0.2024** | 0.4721 |
| $\upsilon = 6$ | **0.2008** | 0.3709 | 0.2126 | 0.4677 |
| $\upsilon = 8$ | **0.1906** | 0.3669 | 0.2148 | 0.4492 |
| $\upsilon = 10$ | **0.1825** | 0.3596 | 0.2137 | 0.4275 |
| $\mathcal{W}(\upsilon, \lambda)$ | | | | |
| $\upsilon = \lambda = 1/2$ | 0.6775 | **0.2729** | 0.3886 | 0.9070 |
| $\upsilon = \lambda = 2$ | 0.1750 | **0.0664** | 0.0938 | 0.4333 |

**Table 7**: MAREs associated with the four estimators of the extreme quantile $q(\alpha_n = 1/n^4)$ computed on simulated data from classical Weibull tail-distributions. The best result is emphasized in bold.

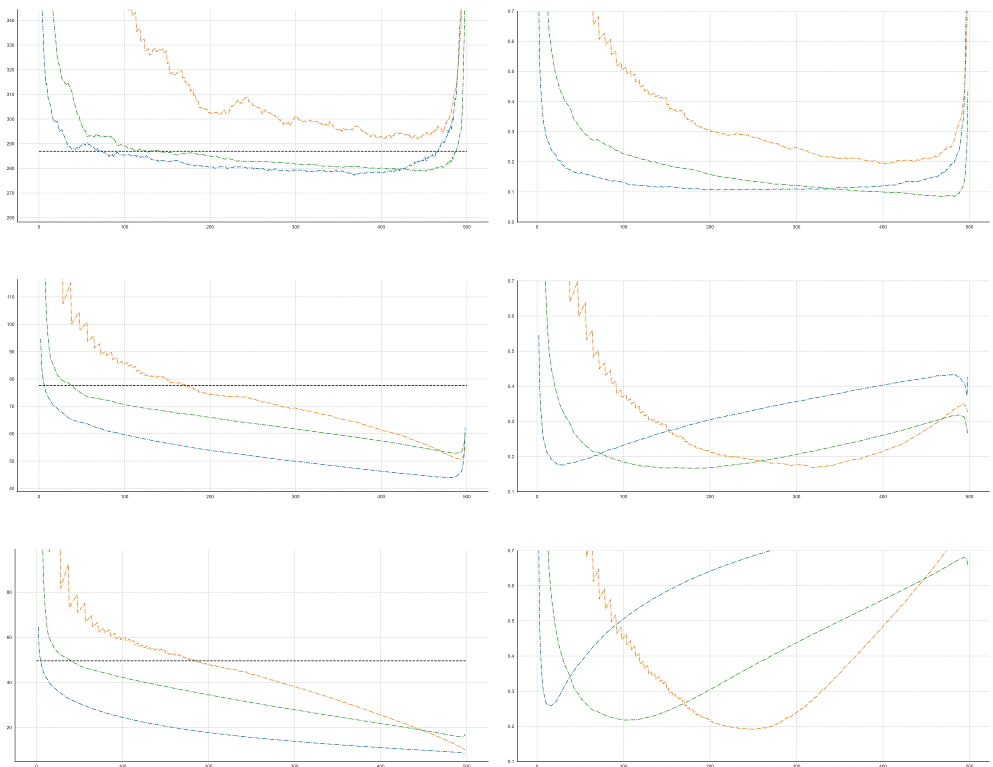

**Figure 2**: Illustration on simulated data sets of size $n = 500$ from a $\mathcal{D}(\zeta = 1/\theta = 2/3, \eta = -\rho, a = 10)$-distribution with $\rho \in \{-1/2, -1, -3/2\}$ (from top to bottom) computed on $N = 1000$ replications. Medians (left panel) and MAREs (right panel) as functions of $k_n \in \{2, \dots, n-1\}$, associated with estimators RWT (orange), RSH (blue) and LSE (green) of the extreme quantile $q(\alpha_n = 1/n^2)$ (black dashed line).

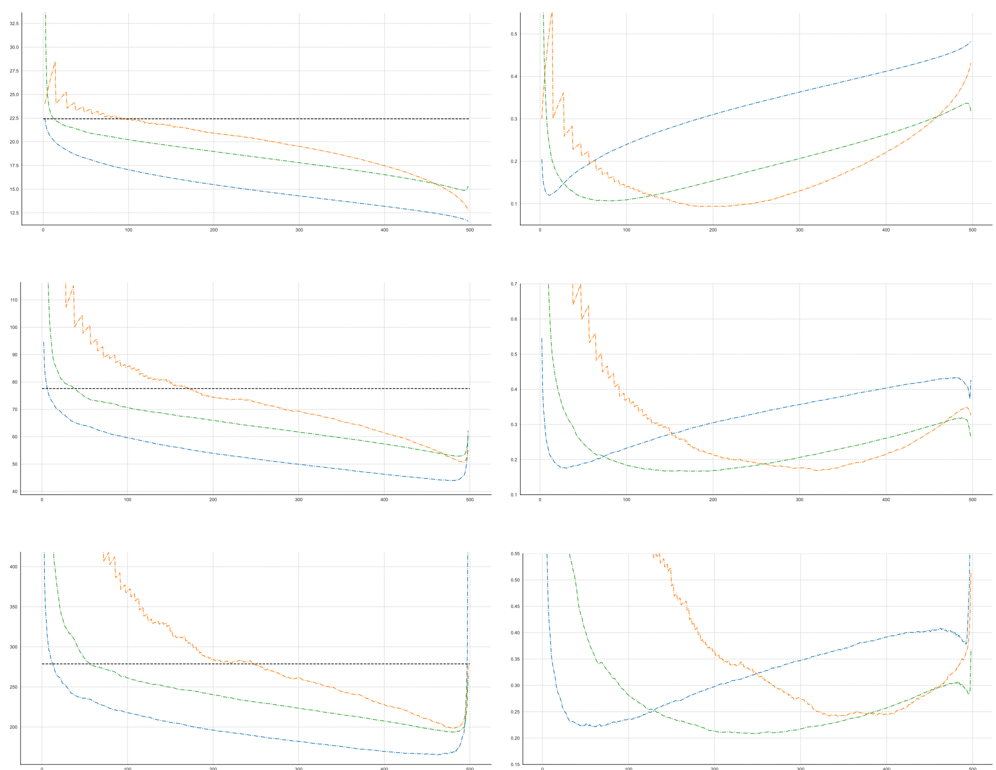

**Figure 3**: Illustration on simulated data sets of size $n = 500$ from a $\mathcal{D}(\zeta = 1/\theta, \eta = -\rho = 1, a = 10)$-distribution with $\theta \in \{1, 3/2, 2\}$ (from top to bottom) computed on $N = 1000$ replications. Medians (left panel) and MAREs (right panel) as functions of $k_n \in \{2, \ldots, n-1\}$, associated with estimators RWT (orange), RSH (blue) and LSE (green) of the extreme quantile $q(\alpha_n = 1/n^2)$ (black dashed line).

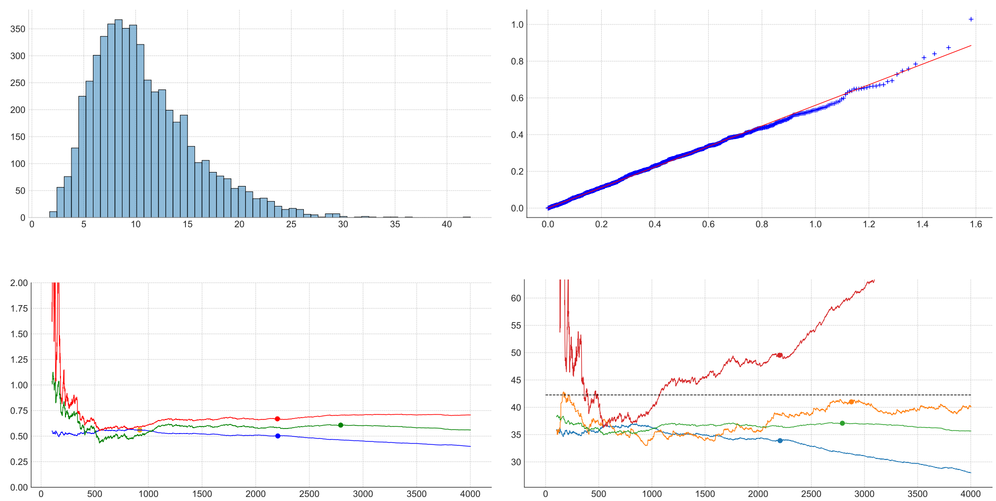


**Figure 4**: Illustration on the daily wind data set. Top-left: Histogram of the data set. Top-right: Weibull quantile-quantile plot (horizontally: $(\log\log(n/i) - \log\log(n/\hat{k}_n^\star))$, vertically: $(\log X_{n-i+1,n} - \log X_{n-\hat{k}_n^\star+1,n})$ for $i \in \{1, \dots, \hat{k}_n^\star = 961\}$). A line with slope $\hat{\theta}_n^{\mathrm{RWT}}(k_n^\dagger) = 0.5597$ is superimposed in red. Bottom-left: Estimates of the Weibull-tail coefficient RSH (blue), LSE (green) and MEF (red) as functions of $k_n$. The range is limited to $k_n \in \{100, \dots, 4000\}$ for the sake ok readability. The pair $(k_n^\dagger, \hat{\theta}_n(k_n^\dagger))$ associated with the selected value $k_n^\dagger$ is emphasized by a circle. The RWT estimate is represented by an orange circle on the RSH curve. Bottom-right: Estimates of the extreme quantile $q(\alpha_n = 1/n)$ by RWT (orange), RSH (blue), LSE (green) and MEF (red) as functions of $k_n$ with their associated $k_n^\dagger$ emphasized by a circle. The sample maximum $x_{n,n}$ is depicted by a black dashed line.

# A. APPENDIX: PROOFS

Proofs of main results are collected in Subsection A.1. Auxiliary results are provided in Subsection A.2 and proved in the Supplementary material document.

## A.1. Proofs of main results

**Proof of Theorem 2.1.** Clearly, the following expansion holds

$$\sqrt{k_n'}\log\left(\frac{\hat{q}_n(\alpha_n,k_n,k_n')}{q(\alpha_n)}\right)=T_n^{(1)}+T_n^{(2)}+T_n^{(3)},$$

$$\text{with } T_n^{(1)} = \sqrt{k_n'}\left(\frac{X_{n-k_n+1,n}}{V(\log(n/k_n))}\right),\ T_n^{(2)}=\sqrt{k_n'}(\log\tau_n)(\hat{\theta}_n^{\mathrm{RSH}}(k_n')-\theta),$$

$$\text{and } T_n^{(3)} = \sqrt{k_n'}\left(\frac{\ell(\log(n/k_n))}{\ell(\log(1/\alpha_n))}\right).$$

Let us consider the three terms separately. First, [16, Lemma 1] shows that, under **(A.1)**, $k_n\to\infty$, $k_n/n\to 0$ and condition (iii):

$$T_n^{(1)}=\frac{\sqrt{k_n'/k_n}}{\log(n/k_n)}\theta\xi_n'+O_P\left(\frac{\sqrt{k_n'}}{k_n\log^2(n/k_n)}\right), \tag{1.1}$$

where $\xi_n'\xrightarrow{d}\mathcal{N}(0,1)$. Second, [17, Proposition 2.1] entails that, under assumptions **(A.1)**, **(A.2)**, $k_n'\to\infty$ and $k_n'/n\to 0$, the following expansion holds

$$T_n^{(2)}=\theta\log(\tau)\xi_n+\theta\log(\tau)\mu(\log(n/k_n'))\xi_n''+\log(\tau)\sqrt{k_n'}b(\log(n/k_n'))(1+o(1)), \tag{1.2}$$

where $\xi_n\xrightarrow{d}\mathcal{N}(0,1)$, $\xi_n''\xrightarrow{d}\mathcal{N}(0,1)$. Third, **(A.2)** and condition (iii) imply

$$T_n^{(3)}=-K_\rho(\tau)\sqrt{k_n'}b(\log(n/k_n))(1+o(1)). \tag{1.3}$$

Collecting (1.1), (1.2) and (1.3), one has

$$\begin{aligned}\sqrt{k_n'}\log&\left(\frac{\hat{q}_n(\alpha_n,k_n,k_n')}{q(\alpha_n)}\right)\\ =&\ \theta\log(\tau)\xi_n+\theta\log(\tau)\mu(\log(n/k_n'))\xi_n''+\frac{\sqrt{k_n'/k_n}}{\log(n/k_n)}\theta\xi_n'+O_P\left(\frac{\sqrt{k_n'}}{k_n\log^2(n/k_n)}\right)\\ +&\ \sqrt{k_n'}\left\{\log(\tau)b(\log(n/k_n'))(1+o(1))-K_\rho(\tau)b(\log(n/k_n))(1+o(1))\right\}.\end{aligned}$$

Recalling that, from [17, Lemma 5.3], $\mu(t)\sim 1/t$ as $t\to\infty$, the above expansion can be simplified as

$$\begin{aligned}\sqrt{k_n'}\log&\left(\frac{\hat{q}_n(\alpha_n,k_n,k_n')}{q(\alpha_n)}\right)\\ =&\ \theta\log(\tau)\xi_n+\frac{\sqrt{k_n'/k_n}}{\log(n/k_n)}\theta\xi_n'+O_P\left(\frac{\sqrt{k_n'}}{k_n\log^2(n/k_n)}\right)\\ +&\ \sqrt{k_n'}\left\{\log(\tau)b(\log(n/k_n'))(1+o(1))-K_\rho(\tau)b(\log(n/k_n))(1+o(1))\right\}.\end{aligned}$$

Finally, remark that assumption (ii) implies $k'_n \le k_n$ eventually and $b(\log(n/k_n)) \sim \beta^{-\rho} b(\log(n/k'_n))$ as $n \to \infty$ so that

$$\sqrt{k'_n} \log\left(\frac{\hat{q}_n(\alpha_n, k_n, k'_n)}{q(\alpha_n)}\right) = \theta \log(\tau)\xi_n + \sqrt{k'_n} b(\log(n/k'_n)) \left(\log(\tau) - \beta^{-\rho} K_\rho(\tau) + o(1)\right)(1 + o(1)) + o_P(1).$$

Assumption (i) then yields

$$\sqrt{k'_n} \log\left(\frac{\hat{q}_n(\alpha_n, k_n, k'_n)}{q(\alpha_n)}\right) \xrightarrow{d} \mathcal{N}\left(\lambda(\log\tau - \beta^{-\rho} K_\rho(\tau)), (\theta \log\tau)^2\right)$$

and a first order Taylor expansion proves the result. □

**Proof of Lemma 2.1:** **(i)** Remarking that $K_y(\tau) \sim -1/y$ as $y \to -\infty$ for all $\tau > 1$ yields $\beta^\star(y, \tau) \to 1$ as $y \to -\infty$. The result $\beta^\star(\tau, 0) := \sqrt{\tau}$ follows from a second-order Taylor expansion.
**(ii)** First, Lemma 1.3(iii) implies that, for all $\tau > 1$ and $y < 0$, $h_y(\tau) > 1$ and thus $\beta^\star(\tau, y) = h_y(\tau)^{-1/y} > 1$. Second, Lemma 1.3(iii) implies that, for all $\tau > 1$ and $y < 0$, $h_y(\tau) < \tau^{-y/2} < \tau^{-y}$ when $y < 0$. This straightforwardly implies that $K_y(\tau)/\log(\tau) > \tau^y$ which is equivalent to $\beta^\star(\tau, y) < \tau$. In the particular case where $y = 0$, from (i), one can take $\beta^\star(\tau, 0) := \sqrt{\tau} < \tau$ since $\tau > 1$ and the result is proved.
**(iii)** Let us first consider $\tilde{k}_n(\tau, y) = n(k_n/n)^{\beta^\star(\tau,y)}$ such that $k^\star_n(\tau, y) = \lfloor \tilde{k}_n(\tau, y) \rfloor$. Clearly, $\tilde{k}_n(\tau, y)/k_n = (k_n/n)^{\beta^\star(\tau,y)-1}$ and $\beta^\star(\tau, y) > 1$ in view of (ii). As a consequence, for all $y \le 0$, $\tilde{k}_n(\tau, y)$ is an increasing function of $k_n$, $\tilde{k}_n(\tau, y) \le k_n$ and $\tilde{k}_n(\tau, y)/k_n \to 0$ as $n \to \infty$. These properties can be extended to $k^\star_n(\tau, y)$ without difficulty since the integer part is an increasing function and $k^\star_n(\tau, y) \le \tilde{k}_n(\tau, y)$.
**(iv)** Routine calculations give for all $\tau > 1$ and $y < 0$,

$$\begin{aligned}\frac{\partial}{\partial y} &\log(\beta^*(\tau, y)) \\ &= \frac{1}{y^2(\tau^y - 1)}\left(\tau^y \log(\tau^y) - \tau^y + 1 - \tau^y \log\left(\frac{\tau^y - 1}{\log(\tau^y)}\right) + \log\left(\frac{\tau^y - 1}{\log(\tau^y)}\right)\right) \\ &=: \frac{1}{y^2(\tau^y - 1)}\varphi(\tau, y).\end{aligned}$$

Letting $x := \tau^y \in (0, 1)$ yields

$$\tilde{\varphi}(x) := \varphi(\tau, y) = x\log(x) - x + 1 - x\log\left(\frac{x - 1}{\log(x)}\right) + \log\left(\frac{x - 1}{\log(x)}\right),$$

and differentiating, one gets

$$\tilde{\varphi}'(x) = -\log\left(\left(1 - \frac{1}{x}\right)\frac{1}{\log(x)}\right) + \left(1 - \frac{1}{x}\right)\frac{1}{\log(x)} - 1 = -\log(u(x)) + u(x) - 1,$$

where $u(x) := (1 - 1/x)/\log(x) > 0$. It thus appears that $\tilde{\varphi}'(x) \ge 0$ for all $x \in (0, 1)$ since $-\log(u) + u - 1 \ge 0$ for all $u > 0$. As a consequence, $\tilde{\varphi}(\cdot)$ is an increasing function on $(0,1)$. Moreover, taking account of $\tilde\varphi(x)\to 0$ as $x\to 1^-$ shows that $\tilde\varphi(x)\le 0$ for all $x\in(0,1)$. Finally, $\tau^y-1<0$ and $\varphi(\tau,y)\le 0$ for all $\tau>1$ and $y<0$ imply that $\beta^*(\tau,y)$ is an increasing function of $y$ which in turns shows that $k_n^\star(\tau,y)$ is a decreasing function of $y$. □

**Proof of Corollary 2.1.** **(i)** To prove the convergence in distribution, it is sufficient to show that conditions (i), (ii) and (iii) of Theorem 2.1 hold. Let $y\le 0$ and $\tau>1$. First, one can easily check that $k_n^\star(\tau,y)b^2(\log n)\to\lambda^2$ as $n\to\infty$ and thus $\sqrt{k_n^\star(\tau,y)}b(\log n)\to\lambda$ in view of the sign assumption on $\lambda$. Besides,

$$\log(n/k_n^\star(\tau,y))=\log n+2\log|b(\log n)|-2\log|\lambda|+o(1)\sim\log n, \tag{1.4}$$

since $b(\cdot)$ is regularly-varying so that $b(\log(n/k_n^\star(\tau,y)))\sim b(\log n)$ and thus

$$\sqrt{k_n^\star(\tau,y)}b(\log(n/k_n^\star(\tau,y)))\to\lambda,$$

as $n\to\infty$. Theorem 2.1(i) is thus proved. Second, observe that

$$\tau_n=\frac{\log(1/\alpha_n)}{\log(n/k_n)}=\frac{\log(1/c)}{\log(n/k_n)}+\frac{\tau}{\beta^\star(\tau,\rho)}\frac{\log n}{\log(n/k_n)}$$

and

$$\log(n/k_n)=\frac{1}{\beta^\star(\tau,\rho)}(\log n+2\log|b(\log n)|-2\log|\lambda|)+o(1)\sim\frac{1}{\beta^\star(\tau,\rho)}\log n, \tag{1.5}$$

as $n\to\infty$. It is thus clear that $\tau_n\to\tau$ as $n\to\infty$, which is Theorem 2.1(ii). Third, Theorem 2.1(iii) is a straightforward consequence of (1.4) and (1.5).
**(ii)** Proposition 1.1 concludes the proof. □

## A.2. Auxiliary results

Let us begin with a Lemma that establishes that the strict Weibull distribution belongs to the Gumbel maximum domain of attraction ($\gamma=0$), and more importantly, with a second-order parameter $\psi=0$. This result illustrates why inference on Weibull-tail distributions may be difficult since the situation $\gamma=\psi=0$ is the most complicated one for classical extreme-value estimators. Let us also recall that, in contrast, the second-order Weilbull parameter is $\rho=-\infty$, see [20, Table 1] and therefore strict Weibull distributions are an easy situation for dedicated Weibull-tail estimators.

**Lemma 1.1.** *Suppose $F$ is the cumulative distribution function of a strict Weibull distribution with shape parameter $\upsilon>0$, $\upsilon\neq 1$ and scale parameter*

*$\lambda > 0$. Then, the associated tail quantile function $U(\cdot) := F^{\leftarrow}(1 - 1/\cdot)$ verifies the second-order condition*

$$\frac{1}{A(t)}\left(\frac{U(tx) - U(t)}{a(t)} - K_\gamma(x)\right) \to \int_1^x s^{\gamma-1} K_\psi(s) ds,$$

*as $t \to \infty$, for all $x > 0$, see [22, Equation (3.4.5)], with $\gamma = 0$, $\psi = 0$, $a(t) = (\lambda/\upsilon)(\log t)^{1/\upsilon - 1}$ and $A(t) = (1 - \upsilon)/(\upsilon \log t)$.*

The following three analytical results are used to prove Proposition 1.1 below.

**Lemma 1.2.** *Let us define for all $(u, \beta) \in (0, 1] \times [0, 1/2]$, $g_\beta(u) := \log(u)(1 + u^\beta) - 2(u - 1)$. Then, $\forall \beta \in [0, 1/2]$ one has $g_\beta(u) < 0$ if $u \in (0, 1)$ and $g_\beta(1) = 0$.*

**Lemma 1.3.** *Let us define, for all $\tau > 1$ and $y \leq 0$,*

$$h_y(\tau) := \frac{K_0(\tau)}{K_y(\tau)} = \frac{y \log(\tau)}{\tau^y - 1} \text{ if } y < 0 \text{ and } h_0(\tau) := 1 \text{ otherwise.}$$

*Then,*

**(i)** *$h_y(\cdot)$ can be extended by continuity letting $h_y(1) = 1$ for all $y \leq 0$.*

**(ii)** *$h_y(\cdot)$ is an increasing function on $[1, \infty)$ for all $y < 0$ (and $h_0(\cdot)$ is a constant function).*

**(iii)** *$1 < h_y(\tau) < \tau^{-y/2}$ for all $\tau > 1$ and $y < 0$ (all three quantities coincide at $y = 0$).*

**Lemma 1.4.** *Let us define, for all $\tau > 1$ and $y \leq 0$,*

$$f_y(\tau) := -\frac{\log(h_y(\tau))}{y \log(\tau)} \text{ if } y < 0 \text{ and } f_0(\tau) = 1/2 \text{ otherwise.}$$

*Then,*

**(i)** *$f_y(\cdot)$ can be extended by continuity letting $f_y(1) = 1/2$ for all $y \leq 0$.*

**(ii)** *$f_y(\cdot)$ is a decreasing function on $[1, \infty)$ for all $y < 0$ ($f_0(\cdot)$ is a constant function).*

**(iii)** *$0 < f_y(\tau) < 1/2$ for all $\tau \geq 1$ and $y < 0$ (the last two quantities coincide at $y = 0$).*

The next Proposition establishes two unexpected results. First, the bias associated with the refined estimator of extreme quantiles is strictly smaller than the bias associated with the original one, even though a misspecification of the second-order Weibull parameter is used. Second, both (asymptotic) biases vanish at $\rho = 0$.

**Proposition 1.1.** *For all $\tau > 1$, $y \leq 0$ and $\rho < 0$, $|B(\beta^\star(\tau, y), \tau, \rho)| < B(1, \tau, \rho)$, see (2.6) and (2.10). Besides, $B(\beta^\star(\tau, y), \tau, 0) = B(1, \tau, 0) = 0$ for all $\tau > 1$ and $y \leq 0$.*

## ACKNOWLEDGMENTS

This work is supported by the French National Research Agency (ANR) in the framework of the Investissements d'Avenir Program (ANR-15-IDEX-02). S. Girard also acknowledges the support of the Chair Stress Test, Risk Management and Financial Steering, led by the French Ecole Polytechnique and its Foundation and sponsored by BNP Paribas.

## REFERENCES

[1] Aarssen, K. and de Haan, L. (1994). On the maximal life span of humans, *Mathematical Population Studies*, **4**, 4, 259–281.

[2] Abramowitz, M. and Stegun, I.A. (1964). *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, (Vol. 55), US Government printing office.

[3] Albert, C.; Dutfoy, A. and Girard, S. (2020). Asymptotic behavior of the extrapolation error associated with the estimation of extreme quantiles, *Extremes*, **23**, 2, 349–380.

[4] Allouche, M.; El Methni, J. and Girard, S. (2023). A refined Weissman estimator for extreme quantiles, *Extremes*, **26**, 545–572.

[5] Asimit, A.V.; Li, D. and Peng, L. (2010). Pitfalls in using Weibull tailed distributions, *Journal of Statistical Planning and Inference*, **140**, 7, 2018–2024.

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

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

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

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

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

[11] Bingham, N. H.; Goldie, C. M. and Teugels, J.L. (1989). *Regular variation*, Cambridge University press.

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

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

[14] Diebolt, J.; Gardes, L.; Girard, S. and Guillou, A. (2008). Bias–reduced estimators of the Weibull tail-coefficient, *Test*, **17**, 2, 311–331.

[15] Diebolt, J.; Gardes, L.; Girard, S. and Guillou, A. (2008). Bias–reduced extreme quantile estimators of Weibull tail-distributions, *Journal of Statistical Planning and Inference*, **138**, 5, 1389–1401.

[16] Gardes, L. and Girard, S. (2005). Estimating extreme quantiles of Weibull tail distributions, *Communications in Statistics–Theory and Methods*, **34**, 5, 1065–1080.

[17] Gardes, L. and Girard, S. (2006). Comparison of Weibull tail-coe?icient estimators, *REVSTAT–Statistical Journal*, **4**, 2, 163–188.

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

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

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

[21] Goegebeur, Y. and Guillou, A. (2011). A weighted mean excess function approach to the estimation of Weibull-type tails, *Test*, **20**, 1, 138–162.

[22] de Haan, L. and Ferreira, A. (2007). *Extreme value theory: an introduction*, Springer Science and Business Media.

[23] Hall, P. (1982). On some simple estimates of an exponent of regular variation, *Journal of the Royal Statistical Society: Series B*, **44**, 1, 37–42.

[24] Hall, P. and Welsh, A.W. (1985). Adaptive estimates of parameters of regular variation, *The Annals of Statistics*, **13**, 331–341.

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

[26] Vladimirova, M.; Girard, S.; Hien, N. and Arbel, J. (2020). Sub-Weibull distributions: generalizing sub-Gaussian and sub-Exponential properties to heavier-tailed distributions, *Stat*, **9**, e318.

[27] Weissman, I. (1978). Estimation of parameters and large quantiles based on the $k$ largest observations, *Journal of the American Statistical Association*, **73**, 364, 812–815.

[28] Worms, J. and Worms, R. (2019). Estimation of extremes for Weibull-tail distributions in the presence of random censoring, *Extremes*, **22**, 667–704.