Statistica Sinica **25** (2015), 1133-1144
doi:http://dx.doi.org/10.5705/ss.2013.272

ADAPTIVE AND MINIMAX OPTIMAL ESTIMATION OF THE TAIL COEFFICIENT

Alexandra Carpentier and Arlene K. H. Kim

University of Cambridge

Abstract: We consider the problem of estimating the tail index α of a distribution satisfying a (α, β) second-order Pareto-type condition, where β is the second-order coefficient. When β is available, it was previously proved that α can be estimated with the optimal rate $n^{-\beta/(2\beta+1)}$. When β is not available, estimating α with the optimal rate is challenging ; so additional assumptions that imply the estimability of β are usually made. We propose an adaptive estimator of α, and show that this estimator attains the rate $\big(n/\log\log n\big)^{-\beta/(2\beta+1)}$ without a priori knowledge of β or additional assumptions. Moreover, we prove that a $\big(\log\log n\big)^{\beta/(2\beta+1)}$ factor is unavoidable by obtaining the companion lower bound.

Key words and phrases: Adaptive estimation, extreme value index, minimax optimal bounds, Pareto-type distributions.

1. Introduction

We consider the problem of estimating the tail index α of an (α, β) second-order Pareto distribution F, given n i.i.d. observations $X_1, \dots, X_n$. We assume that for some $\alpha, \beta, C, C' > 0$,

$$\big|1 - F(x) - Cx^{-\alpha}\big| \leq C' x^{-\alpha(1+\beta)}, \tag{1.1}$$

and write $\mathcal{S}(\alpha, \beta) := \mathcal{S}(\alpha, \beta, C, C')$ for the set of distributions that satisfy this property. Here the tail index α characterizes the heaviness of the tail, and β represents the proximity between F and an α-Pareto distribution $F_\alpha^P : x \in [C^{1/\alpha}, \infty) \to 1 - Cx^{-\alpha}$.

There is an abundant literature on the problem of estimating α. A popular estimator is Hill's estimator (Hill (1975)) (see also Pickands' estimator (Pickands (1975)). Hill (1975) considered α-Pareto distribution for the tail, and suggested an estimator $\hat{\alpha}_H(r)$ of the tail index α based on the order statistics $X_{(1)} \leq \cdots \leq X_{(n)}$ where r is the fraction of order statistics from the tail,

$$\hat{\alpha}_H(r) = \left(\frac{1}{\lfloor rn \rfloor} \sum_{i=1}^{\lfloor rn \rfloor} \frac{\log(X_{(n-i+1)})}{\log(X_{(n-\lfloor rn \rfloor+1)})} \right)^{-1}. \tag{1.2}$$

For more details, see e.g., de Haan and Ferreira (2006).

The limiting distribution of Hill's estimator was found by Hall (1982) when β is known. Under a model that is quite similar to (1.1), he proved that if $rn^{1/(2\beta+1)} \to 0$ as $n \to \infty$, $\sqrt{nr}(\hat{\alpha}_H(r) - \alpha)$ converges in distribution to $N(0, \alpha^2)$. He also considered a more restricted condition (the exact Hall condition, say)

$$\left|1 - F(x) - Cx^{-\alpha}\right| = C'' x^{-\alpha(1+\beta)} + o(x^{-\alpha(1+\beta)}). \tag{1.3}$$

Under (1.3), with the choice of the sample fraction $r^* = Cn^{-1/(2\beta+1)}$ for some constant C, Theorem 2 of Hall (1982) states that $n^{\beta/(2\beta+1)}(\hat{\alpha}_H(r^*) - \alpha)$ converges to a Gaussian distribution with finite mean and variance, depending on the parameters of the true distribution.

The companion lower bound $n^{-\beta/(2\beta+1)}$, under (1.1), was proved by Hall and Welsh (1984). Drees (2001) improved this result by obtaining sharp asymptotic minimax bounds again when β is available. From these results, we know that the second-order parameter β is crucial to understanding the behaviour of the distribution. Indeed, it determines the rate of estimation of α as well as the optimal sample fraction.

In general, β is unknown. To cope with this problem, Hall and Welsh (1985) showed that, under (1.3), it is possible to estimate β in a consistent way, and thus also to estimate the sample fraction r^* consistently by $\hat{r}$. Hall and Welsh (1985) deduced from these results that $\hat{\alpha}_H(\hat{r})$ is asymptotically as efficient as $\hat{\alpha}_H(r^*)$: $n^{\beta/(2\beta+1)}(\hat{\alpha}_H(\hat{r}) - \alpha)$ converges to a Gaussian distribution with the same mean and variance as those resulting from the choice r^*. Their result is pointwise, but not uniform under (1.3), as opposed to the uniform convergence when β is known.

This first result on adaptive estimation was extended in several ways. Gomes et al. (2008) provided more precise ways to reduce the bias of the estimate of α using the estimate of β by supposing a third order condition, and adaptive estimates of α under the third order condition were considered in Gomes et al. (2012). Other methods for estimating r^* have been proposed, e.g., bootstrap (Danielsson et al. (2001)) and regression (Beirlant et al. (1996)). Drees and Kaufmann (1998) considered a method related to Lepski's method (see Lepski (1992) for more details in a functional estimation setting) by choosing the sample fraction that balances the squared bias and the variance of the resulting estimate. They proved that Hill's estimate computed with this sample fraction is asymptotically as efficient as the oracle estimate if F satisfies a condition that is slightly more restrictive than (1.3). Grama and Spokoiny (2008) consider a more general setting than (1.1), but when they apply their results to the exact Hall model (without little o), their estimator obtains the optimal rate up to a $\log(n)$ factor; this is clearly sub-optimal, Hall and Welsh (1985).

We focus here on deriving results for the setting (1.1). Many common distributions (in particular some distributions with change points in the tail) belong to it, and the construction of the lower bound in Hall and Welsh (1984) was proved for this model. To the best of our knowledge though, either the existing results hold in a more restrictive setting than (1.1), typically in a model that is close to the model (1.3) (see e.g., Hall and Welsh (1985); Beirlant et al. (1996); Drees and Kaufmann (1998); Danielsson et al. (2001); Gomes et al. (2008, 2012)), or the convergence rates for (1.1) are worse than one could expect (see e.g., Grama and Spokoiny (2008)). The set of distributions at (1.1) is significantly larger than the set of distributions that satisfy (1.3). Adaptive estimation under (1.1) is more involved since the second-order parameter β is not always estimable (even a consistent estimator does not exist for all distributions in this model), and the adaptive procedures based on estimating β or the oracle sample fraction r^* as in Hall and Welsh (1985) or Gomes et al. (2008, 2012) may not work on all the functions satisfying (1.1).

We construct an adaptive estimator $\hat{\alpha}$ of α under (1.1) and prove that $\hat{\alpha}$ converges to α with the rate $(n/\log\log(n))^{-\beta/(2\beta+1)}$. Thus, for an arbitrarily small $\epsilon > 0$, and some arbitrarily large range I_1 for α and $[\beta_1, \infty)$ for β, there exist large constants $D, E > 0$ such that, for any $n > D\log(\log(n)/\epsilon)$,

$$\sup_{\alpha\in I_1,\beta>\beta_1}\sup_{F\in\mathcal{S}(\alpha,\beta)}\mathbb{P}_F\left(|\hat{\alpha}-\alpha| \geq E\left(\frac{n}{\log(\log(n)/\epsilon)}\right)^{-\beta/(2\beta+1)}\right) \leq \epsilon. \tag{1.4}$$

There is an additional $\big(\log\log(n)\big)^{\beta/(2\beta+1)}$ factor in the rate with respect to the oracle rate, as we adapt over β on a set of distributions where β is not estimable. Although we obtain worse rates, we prove the optimality of our adaptive estimator by obtaining a matching lower bound. Indeed, there exists a small enough constant $E' > 0$ such that for any n large enough, and for any estimator $\tilde{\alpha}$,

$$\sup_{\alpha\in I_1,\beta>\beta_1}\sup_{F\in\mathcal{S}(\alpha,\beta)}\mathbb{P}_F\left(|\tilde{\alpha}-\alpha| \geq E'\left(\frac{n}{\log(\log(n))}\right)^{-\beta/(2\beta+1)}\right) \geq \frac{1}{4}.$$

Both lower and upper bounds containing the $(\log\log(n))^{\beta/(2\beta+1)}$ factor are new to the best of our knowledge (we do not provide a tight scaling factor as in the paper by Novak (2013), but our setting is different and their rate does not involve the additional $(\log\log(n))^{\beta/(2\beta+1)}$ factor). The presence of the $\log\log n$ factor is not unusual in adaptive estimation (see Spokoiny (1996) in a signal detection setting). This issue is also discussed in Drees and Kaufmann (1998).

The adaptive estimator $\hat{\alpha}$ we propose is based on a sequence of estimates $\hat{\alpha}(k)$, defined in (3.1), where the parameter $k \in \mathbb{N}$ plays a role similar to the

sample fraction in Hill's estimator (see Subsection 3.1 for more details). These estimates $\hat{\alpha}(k)$ are based not on order statistics, but on probabilities of tail events. We first prove that for an appropriate choice of this threshold k (independent of α or β), $\hat{\alpha}(k)$ is consistent. We then prove that for an oracle choice of k (as a function of β), this estimate is minimax-optimal for distributions satisfying (1.1) with the rate $n^{-\beta/(2\beta+1)}$. An adaptive version of this estimate, where the parameter k is chosen in a data-driven way without knowing β in advance, is shown to satisfy (1.4). All the proofs for the results provided in this paper are in the Supplementary Material.

2. Definitions of Distribution Classes

In this section, we introduce the class of approximately α-Pareto distributions, and the class of approximately (α, β) second-order Pareto distributions. We let $\mathcal{D}$ be the class of distribution functions on $[0, \infty)$.

Definition 1. For $\alpha > 0$, $C > 0$, the class of approximately α-Pareto distributions is

$$\mathcal{A}(\alpha, C) = \Big\{ F \in \mathcal{D} : \lim_{x \to \infty} (1 - F(x)) x^{\alpha} = C \Big\}.$$

Distributions in $\mathcal{A}(\alpha, C)$ converge to Pareto distributions for large x, and these distributions have been used as a first attempt to understand heavy tail behavior (see Hill, 1975; de Haan and Ferreira, 2006). The first-order parameter α characterizes the tail behavior in that distributions with smaller α correspond to heavier tails.

Definition 2. For $\alpha > 0$, $C > 0$, $\beta > 0$, and $C' > 0$, the class of approximately (α, β) second-order Pareto distributions is

$$\mathcal{S}(\alpha, \beta, C, C') = \Big\{ F \in \mathcal{D} : \forall x \text{ s.t. } F(x) \in (0, 1], \big|1 - F(x) - C x^{-\alpha}\big| \le C' x^{-\alpha(1+\beta)} \Big\}. \tag{2.1}$$

The rate of approximation here is linked to the second-order parameter β—a large β corresponds to a distribution that is close to a Pareto distribution (in particular, when $\beta = \infty$, it is Pareto), and a small β corresponds to a distribution that is well approximated by a Pareto distribution only for large x. When there is no confusion, we call the distributions in $\mathcal{S}(\alpha, \beta, C, C')$ second-order Pareto distributions, and use the notation $\mathcal{A}$ and $\mathcal{S}$ without writing parameters explicitly.

The condition (2.1) is weaker than the condition (1.3), for (1.3) implies

$$\lim_{x \to \infty} \frac{1 - F(x) - C x^{-\alpha}}{x^{-\alpha(1+\beta)}} = C',$$

whereas our condition imposes only an upper bound,

$$\limsup_{x\to\infty}\Big|\frac{1-F(x)-Cx^{-\alpha}}{x^{-\alpha(1+\beta)}}\Big|\leq C'.$$

This difference is essential in the estimation problem : under (1.3), it is possible to estimate β consistently (see e.g., Hall and Welsh (1985)), whereas under (2.1), it is not possible to estimate β consistently over the set $\mathcal{S}$ of distributions for $\beta\in[\beta_1,\beta_2]$ with $0<\beta_1<\beta_2$. Adaptive estimation of α is thus likely to be more involved in our setting than in under (1.3). Many adaptive techniques rely on estimating β or the sample fraction as a function of β, and are not directly applicable in our setting (see e.g., Hall and Welsh (1985); Danielsson et al. (2001); Gomes et al. (2012)).

Remark 1. The difference between the functions satisfying 2.1 and (1.3) is related to the difference between Hölder functions that actually attain their Hölder exponent and Hölder functions that are in a given Hölder ball but do not attain their Hölder exponent (see e.g., Giné and Nickl (2010) for a comparison of these two sets, and the problem for estimation when the second set is considered).

3. Main Results

Most estimates in the literature are based on order statistics and are difficult to analyse in a non-asymptotic way, while our estimate is based on probabilities of well chosen tail events.

3.1. A new estimate

Let $X_1,\dots,X_n$ be an i.i.d. random sample from a distribution $F\in\mathcal{A}$. We write, for any $k\in\mathbb{N}$, $p_k:=\mathbb{P}(X>e^k)=1-F(e^k)$, and its empirical estimate as $\hat{p}_k:=(1/n)\sum_{i=1}^n\mathbf{1}\{X_i>e^k\}$. For any $k\in\mathbb{N}$, let

$$\hat{\alpha}(k):=\log(\hat{p}_k)-\log(\hat{p}_{k+1}). \tag{3.1}$$

Lemma 1 (Large deviation inequality)**.** *Let $X_1,\dots,X_n$ be an i.i.d. sample from F.*

A. *Suppose $F\in\mathcal{A}$ and let $\delta>0$. For any k such that $p_{k+1}\geq 16\log(2/\delta)/n$, with probability larger than $1-2\delta$,*

$$\big|\hat{\alpha}(k)-(\log(p_k)-\log(p_{k+1}))\,\big|\leq 6\sqrt{\frac{\log(2/\delta)}{np_{k+1}}}. \tag{3.2}$$

B. *Suppose $F \in \mathcal{S}$ and let $\delta > 0$. For any k such that $p_{k+1} \geq 16\log(2/\delta)/n$ and $e^{-k\alpha\beta} \leq C/(2C')$, with probability larger than $1 - 2\delta$,*

$$\left|\hat{\alpha}(k) - \alpha\right| \leq 6\sqrt{\frac{\log(2/\delta)}{np_{k+1}}} + \frac{3C'}{C}e^{-k\alpha\beta} \tag{3.3}$$

$$\leq 6\sqrt{\frac{e^{(k+1)\alpha+1}\log(2/\delta)}{Cn}} + \frac{3C'}{C}e^{-k\alpha\beta}. \tag{3.4}$$

The proof of this lemma is in the Supplementary Material (see Section S2).

For $\hat{\alpha}(k)$, k plays a similar role as the sample fraction in Hill's estimate (1.2). The bias-variance trade-off is solved by choosing k in an appropriate way as a function of β.

3.2. Rates of convergence

For the set of approximately Pareto distributions, we prove that the estimate $\hat{\alpha}(k_n)$ is consistent if we choose k_n diverging to ∞ but not too quickly.

Theorem 1 (Consistency in $\mathcal{A}$)**.** *For $F \in \mathcal{A}$ and $k_n \to \infty$ with $(\log(n)/n)e^{k_n\alpha} \to 0$ as $n \to \infty$, then $\hat{\alpha}(k_n) \to \alpha$ a.s.*

The proof of this theorem is in the Supplementary Material (see Section S3).

The estimate $\hat{\alpha}(\log\log(n))$ converges to α almost surely under the rather weak assumption that F belongs to $\mathcal{A}$. But on such sets, no uniform rate of convergence exists, and so the restricted set $\mathcal{S}$ is introduced.

Let $\alpha, \beta, C, C' > 0$. Consider now the set $\mathcal{S} := \mathcal{S}(\alpha, \beta, C, C')$ of second-order Pareto distributions. We assume to begin with that, although we do not have access to α, we know $\alpha(2\beta + 1)$. This is not realistic, but we can modify the estimate so that it is minimax optimal on the class of second-order Pareto distributions.

Theorem 2 ($\alpha(2\beta + 1)$ is known)**.** *Let n be such that* (S4.1) *is satisfied. Let $k_n^* = \lfloor \log(n^{1/\alpha(2\beta+1)}) + 1 \rfloor$. Then for any $\delta > 0$, we have*

$$\sup_{F \in \mathcal{S}} \mathbb{P}_F\left(|\hat{\alpha}(k_n^*) - \alpha| \geq \left(B_1 + \frac{3C'}{C}\right) n^{-\beta/(2\beta+1)}\right) \leq 2\delta,$$

where $B_1 = 6\sqrt{e^{2\alpha+1}[\log(2/\delta)/C]}$.

The proof of this theorem is in the Supplementary Material (see Section S4).

Theorem 2 states that, uniformly on the class of second-order Pareto distributions, the estimate $\hat{\alpha}(k_n^*)$ converges to α with the minimax optimal rate $n^{-\beta/(2\beta+1)}$ (see Hall and Welsh (1984) for the matching lower bound).

Remark 2. Theorem 2 can be used to prove the convergence rate of our estimator by modifying the choice of k_n^*, when $\alpha(2\beta+1)$ is unknown but only β is known. For instance, we can plug a rough estimate $\tilde{\alpha} := \hat{\alpha}((\log\log(n))^2)$ of α into k_n^*. The idea behind this choice is that with sufficiently large n, we have with high probability,

$$|\hat{\alpha}((\log\log(n))^2) - \alpha| = O\left(\frac{1}{\log n}\right).$$

Then $\hat{k}_n^1$ is defined as $\lfloor \log(n^{1/[\tilde{\alpha}(2\beta+1)]}) + 1 \rfloor$. Finally, the rate of convergence of $\hat{\alpha}(\hat{k}_n^1)$ can be shown as $n^{-\beta/(2\beta+1)}$ by proving $\exp(\hat{k}_n^1) = O(n^{1/(\alpha(2\beta+1))})$ with high probability.

However, the previous optimal choice of k (k_n^* or $\hat{k}_n^1$) still depends on β, which is unavailable in general. To deal with this problem, we construct an adaptive estimate of α that does not depend on β but still attains a rate that is quite close to the minimax optimal rate $n^{-\beta/(2\beta+1)}$ on the class of β second-order Pareto distributions.

The adaptive estimator is obtained by considering a kind of bias and variance trade-off based on the large deviation inequality (3.2). Suppose we know the optimal choice of k^*. Then this k^* optimizes the squared error by making bias and standard error (of the estimate with respect to its expectation) equal. Since the bias is decreasing while the standard error is increasing as k increases, for all k' larger than this optimal k^*, the bias is smaller than the standard error. Based on this heuristic (originally proposed by Lepski (1992)), we pick the smallest k which satisfies, for all k' larger than k, the proxy for the bias is smaller than the proxy for the standard error $O(\sqrt{1/(n\hat{p}_{k'+1})})$, as in (3.2). For the proxy for the bias, we use $|\hat{\alpha}(k') - \hat{\alpha}(k)|$ by treating $\hat{\alpha}(k)$ as the true α based on the idea that $\hat{\alpha}(k)$ would be close in terms of the rate to the true α (if k is selected in an optimal way).

More precisely, we choose k as, for $1/4 > \delta > 0$

$$\hat{k}_n = \inf\left\{ k \in \mathbb{N} : \ \hat{p}_{k+1} > \frac{24\log(2/\delta)}{n} \ \text{ and } \right.$$
$$\left. \forall k' > k \ \text{ s.t. } \hat{p}_{k'+1} > \frac{24\log(2/\delta)}{n}, \ \ |\hat{\alpha}(k') - \hat{\alpha}(k)| \le A(\delta)\sqrt{\frac{1}{n\hat{p}_{k'+1}}} \right\}, \quad (3.5)$$

where $A(\delta)$ satisfies (3.6) below.

Theorem 3 (Rates of convergence with unknown β)**.** *Let $1/4 > \delta > 0$ and let n be such that* (S.5.2) *is satisfied. Consider the adaptive estimator $\hat{\alpha}(k_n)$ where k_n*

is chosen as at (3.5) *where* $A(\delta)$ *satisfies*

$$A(\delta) \geq 6\sqrt{2(C+C')\log(\frac{2}{\delta})}\left(2\sqrt{\frac{e^{2\alpha+1}}{C}}+\frac{C'}{C}\right). \tag{3.6}$$

Then we have

$$\sup_{F\in\mathcal{S}} \mathbb{P}_F\left(|\hat{\alpha}(\hat{k}_n)-\alpha| \geq \left(B_2+\frac{3C'}{C}\right)\left(\frac{n}{\log(2/\delta)}\right)^{-\beta/(2\beta+1)}\right)$$
$$\leq \left(1+\frac{1}{\alpha}\log\left(\frac{(C+C')n}{16}\right)\right)\delta,$$

where $B_2 = \left(B_1+2A(\delta)\sqrt{e^{2\alpha}/C}\right)[1/\sqrt{\log(2/\delta)}]$ *and* B_1 *is defined in Theorem* 2.

The proof of this theorem is in the Supplementary Material (see Section S5).

Theorem 3 holds for any (α,β) provided that n and $A(\delta)$ are larger than some constants depending on α,β,C,C', and on the probability δ. The advantage of our adaptive estimator is that since the threshold $\hat{k}_n$ is chosen adaptively to the samples, the second-order parameter β does not need to be known in the procedure in order to obtain the convergence rate of $\hat{\alpha}(\hat{k}_n)$.

Corollary 1. *Let* $\epsilon \in (0,1)$ *and* $C'>0$, *and let* $0<\alpha_1<\alpha_2$ *and* $0<C_1<C_2$. *We use* $\hat{k}_n$ *as in* (3.5) *where* $A(\delta)=A(\delta(\epsilon))=:A(\epsilon)$ *is chosen as in* (S6.1). *If* n *satisfies* (S6.3), *then*

$$\sup_{\substack{\alpha\in[\alpha_1,\alpha_2],\beta\in[\beta_1,\infty]\\ C\in[C_1,C_2]}} \sup_{F\in\mathcal{S}(\alpha,\beta,C,C')} \mathbb{P}_F\Bigg(|\hat{\alpha}(\hat{k}_n)-\alpha|$$
$$\geq B_3\left(\frac{n}{\log\left((2/\epsilon)\left(1+\log((C_2+C')n)/\alpha_1\right)\right)}\right)^{-\beta/(2\beta+1)}\Bigg) \leq \epsilon,$$

where B_3 *is a constant explicitly expressed in* (S6.2), *which only depends on* α_2, C_1, C_2, *and* C'.

The proof of this corollary is in the Supplementary Material (see Section S6).

In other words, if we fix the range of the α and C and a lower bound on β to which we wish to adapt, we can tune the parameters of the adaptive choice of $\hat{k}_n$ so that we adapt to the maximal β such that F is β second-order Pareto. Moreover, this adaptive procedure works uniformly well over the set of second-order Pareto distributions satisfying (1.1) (for $\alpha\in[\alpha_1,\alpha_2],\beta\in[\beta_1,\infty],C\in[C_1,C_2]$), which is much larger than the class of distributions that verify the condition (1.3). Then this gives *non-asymptotic guarantees* with *explicit bounds*.

Remark 3. The parameter C' plays a role in the definition of the second order Pareto class that is slightly different than the one of C or α, β. Unlike α or C, C' is not uniquely defined: if $F \in \mathcal{S}(\alpha, \beta, C, \tilde{C}')$, then $F \in \mathcal{S}(\alpha, \beta, C, C')$ with $C' \geq \tilde{C}'$. This implies in particular that the results of Corollary 1 could have been rewritten, fixing a constant $C' > 0$ and writing $\tilde{C}'$ for a constant that fits more closely F, by taking supremum over $F \in \mathcal{S}(\alpha, \beta, C, \tilde{C}')$ where $\tilde{C}' \leq C'$. Being non-adaptive over $\tilde{C}'$ and choosing a loose constant C' instead of $\tilde{C}'$ will only worsen the bound by a constant factor, unlike making a mistake on β which will worsen the exponent of the bound.

It seems that we lose a $(\log\log(n))^{\beta/(2\beta+1)}$ factor with respect to the optimal rate, due to adaptivity to β. However, the lower bound below implies that this $(\log\log(n))^{\beta/(2\beta+1)}$ loss is inevitable; hence the rate provided in Theorem 3 is sharp.

Theorem 4 (Lower bound)**.** *Let $\alpha_1, \beta_1, C_1, C_2, C' > 0$ be such that $C_1 \leq \exp(-1/2\alpha_1(2\beta_1+1))$, $C_2 \geq 1$ and $C' \geq 1/2\alpha_1\beta_1$. Let n be sufficiently large. Then for any estimate $\tilde{\alpha}$ of α,*

$$\sup_{\substack{\alpha\in[\alpha_1,2\alpha_1],\beta\in[\beta_1,\infty)\\ C\in[C_1,C_2]}} \sup_{F\in\mathcal{S}(\alpha,\beta,C,C')} \mathbb{P}_F\left(|\tilde{\alpha}-\alpha| \geq B_4\Big(\frac{n}{\log\big(\log(n)/2\big)}\Big)^{-\beta/(2\beta+1)}\right) \geq \frac{1}{4},$$

where B_4 is a constant depending on α_1 and β_1, which is provided in (S7.9).

The proof of this theorem is in the Supplementary Material (see Section S7).

The lower bound result is proved with specific ranges of the parameters (e.g., restrictions on C_1, C_2, C' in the statement of Theorem 4). but it can be modified by considering different ranges (see Remark 1 in S8 in the Supplementary Material).

3.3. Additional remarks on our estimate

In the definition of our estimate, we use exponential spacings (i.e., we estimate the probability that the random variable is larger than e^k), but we can generalize our estimate by considering the probability of other tail events. For some parameters $u > v \geq 1$, define

$$\hat{q}_u = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_i > u\}, \text{ and } \hat{q}_v = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_i > v\}.$$

We define the following estimate of α as

$$\hat{\alpha}(u, v) = \frac{\log(\hat{q}_v) - \log(\hat{q}_u)}{\log(u) - \log(v)}. \tag{3.7}$$

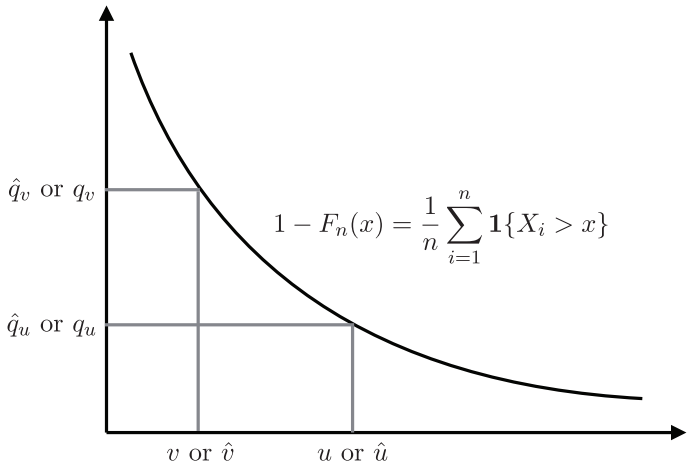

Figure 1. Duality between the estimate (3.7) and the estimate (3.8).

If we fix $v \sim O(n^{1/(\alpha(2\beta+1)})$ and $u/v \sim O(1)$, then we will also obtain the oracle rate for estimating α with $\hat{\alpha}(u, v)$. However, the choice of u/v will have an impact on the constants. In practice, these parameters are important to tune well (in particular for the exact Pareto case, or for distributions satisfying Equation (1.3)). However, a precise analysis of the best choices for u and v (in terms of constants) is beyond the scope of this paper.

Another point we want to address is the relation between our estimate and usual estimates based on order statistics. To estimate the tail index α, it is natural to consider the quantiles associated with the tail probabilities. For the estimates based on order statistics, one fixes some tail-probabilities and then observes the order statistics in order to estimate the quantiles. On the other hand, we fix some values corresponding to the quantiles, and estimate the associated tail probabilities. Based on such a link, one could relate any existing method based on order statistics to the method based on tail probabilities.

In particular, the estimator based on order statistics corresponding to our estimator would be of the form, for some parameters $1 \geq q_v > q_u \geq 0$,

$$\tilde{\alpha}(q_u, q_v) = \frac{\log(q_v) - \log(q_u)}{\log(\hat{u}) - \log(\hat{v})}, \tag{3.8}$$

where $\hat{u} = X_{(n-\lfloor q_u n \rfloor)}$ and $\hat{v} = X_{(n-\lfloor q_v n \rfloor)}$. This estimate can be interpreted as the inverse of some generalized Pickands' estimate (see Pickands (1975), it is however *not* Pickands' estimate). There is actually a duality between these two estimators: for any couple (q_u, q_v) in the definition (3.8), it is possible to find (u, v) in the definition (3.7) such that these two estimates exactly match (see Figure 1 for an illustration). However, there is no analytical transformation from one estimate to the other since such a transformation will be data dependent.

Acknowledgement

We are grateful to Richard J. Samworth and Richard Nickl for their comments and advice. We are also grateful to the referees, an associate editor and the editor for insightful comments that were helpful in enhancing the quality of the paper.

References

Beirlant, J., and Vynckier, P. and Teugels, J. (1996). Tail index estimation, Pareto quantile plots and regression. *J. Amer. Statist. Assoc.* **70**, 1659-1667.

Danielsson, J., de Haan, L., Peng, L. and de Vries, C. G. (2001). Using a bootstrap method to choose the sample fraction in tail index estimation. *J. Multivariate Anal.* **2**, 226-248.

Drees, H. (2001). Minimax risk bounds in extreme value theory. *Ann. Statist.* **29**, 266-294.

Drees, H. and Kaufmann, E. (1998). Selecting the optimal sample fraction in univariate extreme value estimation. *Stochastic Process. Appl.* **75**, 149-172.

Giné, E. and Nickl, R. (2010). Confidence bands in density estimation. *Ann. Statist.* **38**, 1122-1170.

Gomes, M. I., Figueiredo, F. and Neves, M. (2012). Adaptive estimation of heavy right tails: resampling-based methods in action. *Extremes* **15**. 463-489.

Gomes, I. M, De Haan, L. and Rodrigues, L. H. (2008). Tail index estimation for heavy-tailed models: accommodation of bias in weighted log-excesses. *J. Roy. Statist. Soc. Ser. B* **91**, 31-52.

Grama, I. and Spokoiny, V. (2008). Statistics of extremes by oracle estimation. *Ann. Statist.* **36**, 1619-1648.

de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory: An Introduction*. Springer.

Hall, P. (1982). On some simple estimates of an exponent of regular variation. *J. Roy. Statist. Soc. Ser. B* **44**, 37-42.

Hall. P. and Welsh, A. H. (1984). Best attainable rates of convergence for estimates of parameters of regular variation. *Ann. Statist.* **12**, 1079-1084.

Hall. P. and Welsh, A. H. (1985). Adaptive estimates of parameters of regular variation. *Ann. Statist.* **75**, 331-341.

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *Ann. Statist.* **3**, 1163-1174.

Lepski, O. V. (1992). On problems of adaptive estimation in white gaussian noise. *Topics in Nonparametric Estimation* **12**, 87-106.

Novak, S. Y. (2013). Lower bounds to the accuracy of inference on heavy tails. *Bernoulli*. To appear.

Pickands, J. (1975). Statistical inference using extreme order statistics. *Ann. Statist.* **3**, 119-131.

Spokoiny, V. G. (1996). Adaptive hypothesis testing using wavelets. *Ann. Statist.* **24**, 2477-2498.

Statistical Laboratory, Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, Wilberforce Road, CB3 0WB Cambridge, United Kingdom.

E-mail: a.carpentier@statslab.cam.ac.uk

Statistical Laboratory, Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, Wilberforce Road, CB3 0WB Cambridge, United Kingdom.

E-mail: a.kim@statslab.cam.ac.uk

(Received September 2013; accepted November 2014)