Electronic Journal of Statistics
Vol. 13 (2019) 1872–1925
ISSN: 1935-7524
https://doi.org/10.1214/19-EJS1561

Data-adaptive trimming of the Hill estimator and detection of outliers in the extremes of heavy-tailed data

Shrijita Bhattacharya* **and Michael Kallitsis†**

Department of Statistics and Probability, C419 Wells Hall, 619 Red Cedar Rd, East Lansing, MI 48824
Merit Network, 1000 Oakbrook Drive, Suite 200, Ann Arbor, Michigan 48104
e-mail: bhatta61@msu.edu; mgkallit@merit.edu

Stilian Stoev

Department of Statistics, 445C W Hall, 1085 S. University Ann Arbor, MI 48109
e-mail: sstoev@umich.edu
url: https://sites.lsa.umich.edu/sstoev/

Abstract: We introduce a trimmed version of the Hill estimator for the index of a heavy-tailed distribution, which is robust to perturbations in the extreme order statistics. In the ideal Pareto setting, the estimator is essentially finite-sample efficient among all unbiased estimators with a given strict upper break-down point. For general heavy-tailed models, we establish the asymptotic normality of the estimator under second order regular variation conditions and also show that it is minimax rate-optimal in the Hall class of distributions. We also develop an automatic, data-driven method for the choice of the trimming parameter which yields a new type of robust estimator that can *adapt* to the unknown level of contamination in the extremes. This adaptive robustness property makes our estimator particularly appealing and superior to other robust estimators in the setting where the extremes of the data are contaminated. As an important application of the data-driven selection of the trimming parameters, we obtain a methodology for the principled identification of extreme outliers in heavy tailed data. Indeed, the method has been shown to correctly identify the number of outliers in the previously explored Condroz data set.

MSC 2010 subject classifications: Primary 62G32, 62G35; secondary 62G30.
Keywords and phrases: Trimmed Hill, adaptive robustness, weighted sequential testing, minimax rate optimality.

Received August 2018.

Contents

*Supported by National Science Foundation grant CNS-1422078.
†Supported by the NSF grant ATD grant DMS–1830293.

1. Introduction

The estimation of the tail index for heavy-tailed distributions is perhaps one of the most studied problems in extreme value theory. Since the seminal works of [23, 26, 33] among many others, numerous aspects of this problem and its applications have been explored (see e.g., the monographs [7] and [19]).

Let $X_1, \cdots, X_n$ be an i.i.d. sample from a distribution F. We shall say that F has a heavy (right) tail if:

$$\mathbb{P}(X_1 > x) \equiv 1 - F(x) = \ell(x)x^{-1/\xi}, \tag{1.1}$$

for some $\xi > 0$ and a slowly varying function $\ell : (0, \infty) \to (0, \infty)$, i.e., $\ell(\lambda x)/\ell(x) \to 1$, $x \to \infty$, for all $\lambda > 0$. The parameter ξ is referred to as the *tail index* of F. Its estimation is of fundamental importance to the applications of extreme

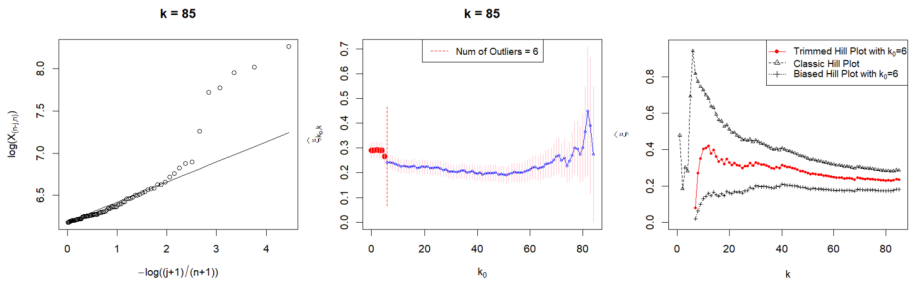

FIG 1. *Exploratory plots of the Condroz data set.* Left: *Pareto quantile plot,* Middle: *Diagnostic Plot and* Right: *Hill plots viz classic Hill plot, trimmed Hill plot and biased Hill plot.*

value theory (see for example the monographs [7], [15], [34], and the references therein).

The fact that the tail index ξ governs the asymptotic right tail-behavior of F means that, in practice, one should estimate it by focusing on the most extreme values of the sample. In many applications, one may quickly run out of data since only the largest few order statistics are utilized. Since every extreme data-point matters, the problem becomes even more challenging when a certain number of these large order statistics are *corrupted*. Contamination of the top order statistics, if not properly accounted for, can lead to severe bias in the estimation of the tail index. For example, the right panel of Figure 1 shows the classic Hill plot, its biased version and our new trimmed Hill plot for the Condroz data set which has been previously identified to have 6 outliers (see [35, 37] and Section 5.1, below, for more details). We shall elaborate more on the construction of these three plots in the rest of the introduction but observe the drastic difference in the tail-index estimates produced by these methods (see also the R shiny app at [29]).

Recall the *classic Hill* estimator of ξ:

$$\widehat{\xi}_k(n) := \frac{1}{k}\sum_{i=1}^{k} \log\left(\frac{X_{(n-i+1,n)}}{X_{(n-k,n)}}\right), \qquad 1 \le k \le n-1. \tag{1.2}$$

It is based on the top-k of the order statistics:

$$X_{(n,n)} \ge X_{(n-1,n)} \ge \cdots \ge X_{(1,n)}$$

of the sample $X_i,\ i = 1, \cdots, n$.

Naturally, one can trim a certain number of the largest order statistics in order to obtain a robust estimator of ξ. This idea has already been considered in Brazauskas and Serfling [12], who (among other robust estimators) defined a trimmed version of the Hill estimator:

$$\widehat{\xi}^{\text{trim}}_{k_0,k}(n) := \sum_{i=k_0+1}^{k} c_{k_0,k}(i) \log\left(\frac{X_{(n-i+1,n)}}{X_{(n-k,n)}}\right), \qquad 0 \le k_0 < k < n. \tag{1.3}$$

where the weights $c_{k_0,k}(i)$ were chosen so that the estimator is asymptotically unbiased for ξ (see Section 3.1 in [12]). The weights used by Brazauskas and Serfling, however, are not optimal. In Section 2.1, we show that the asymptotically optimal trimmed Hill estimator has the form

$$\widehat{\xi}_{k_0,k}(n) := \frac{k_0}{k-k_0}\log\left(\frac{X_{(n-k_0,n)}}{X_{(n-k,n)}}\right) + \underbrace{\frac{1}{k-k_0}\sum_{i=k_0+1}^{k}\log\left(\frac{X_{(n-i+1,n)}}{X_{(n-k,n)}}\right)}_{\widehat{\xi}^{0}_{k_0,k}(n)} \tag{1.4}$$

for $0 \le k_0 < k < n$. Note that if $k_0 = 0$ the *trimmed Hill* estimator $\widehat{\xi}_{k_0,k}(n)$ coincides with the classic Hill estimator.

A number of authors have also considered trimming but of the *models* rather than the data. Specifically, the seminal works of [2] and [6] studied the case where the distribution is *truncated* to a potentially unknown large value. In contrast, here we assume to have non-truncated heavy-tailed model and trim the data as a way of achieving robustness to outliers in the extremes.

Suppose now that somehow one has identified that the top-k_0 order statistics have been corrupted. Following [30], if one were to simply ignore them and apply the classic Hill estimator to the observations $X_{(n-k_0)} \ge \cdots \ge X_{(n-k,n)}$, the estimator would be biased. Indeed, the second summand, $\widehat{\xi}^{0}_{k_0,k}(n)$ in (1.4) gives the expression for this *biased Hill* estimator. The recent work of Zou *et al* [39] uses this biased Hill estimator in a different inferential censoring–type context, where *an unknown* number k_0 of the top order statistics is missing.

Let us return to Figure 1 (right panel) based on the Condroz data set. It shows the classic Hill plot, i.e., the plot of $\widehat{\xi}_k(n)$ as a function of k as well as the plots of $\widehat{\xi}_{k_0,k}(n)$ and $\widehat{\xi}^{0}_{k_0,k}(n)$ as a function of k. We refer to the last two plots as to the trimmed Hill and biased Hill plots, respectively. Since the data exhibits six outliers (Figure 1 left panel), the trimmed Hill and biased Hill plots are based on $k_0 = 6$. The significant difference in the three plots demonstrates the effect that outliers can have on the estimation of the tail index.

In this paper, we introduce and study the trimmed Hill estimator $\widehat{\xi}_{k_0,k}(n)$ defined in (1.4). We begin by establishing its finite sample optimality and robustness properties. Specifically, for ideal Pareto data, we establish in Theorem 2.5 that the trimmed Hill estimator has nearly minimum-variance among all unbiased estimators with a given *strict upper break-down point* (see Definition 2.4). For heavy tailed models in (1.1), since the Pareto regime emerges asymptotically, it is not surprising that the trimmed Hill estimator is also minimax rate-optimal. This was shown in Theorem 3.2 for the Hall class of heavy-tailed distributions. Furthermore, under technical second-order regular variation conditions, we establish the asymptotic normality of the trimmed Hill estimator in Section 3.2.

The optimality and asymptotic properties of the trimmed Hill estimator although interesting are not practically useful unless one has a data-adaptive method for the choice of the trimming parameter k_0. This problem is addressed

in Section 2.2. We start by introducing a *diagnostic plot* to visually determine the number of outliers k_0. It is a plot of the trimmed Hill estimator as function of k_0 for a fixed k. Figure 1 (middle panel) displays this plot for the Condroz data set with previously identified six outliers. A sudden change point at $k_0 = 6$ further corroborates the hypothesis of six plausible outliers in the data set. This value of k_0 was automatically identified by the method we introduce in Section 2.2. The methodology for the automatic selection of k_0 is based on a weighted sequential testing method, which exploits the elegant structure of the joint distribution of $\widehat{\xi}_{k_0,k}(n),\ k_0 = 0, 1, \ldots, k-1$ in the ideal Pareto setting. In Section 3.2, we show that this test is asymptotically consistent in the general heavy-tailed regime (1.1) under second order conditions on the regularly varying function ℓ of [4]. In fact, the resulting estimator $\widehat{\xi}_{\widehat{k}_0,k}(n)$, where $\widehat{k}_0$ is automatically selected, has an excellent finite sample performance and is *adaptively robust* to shifting degrees of contamination in the data. This novel adaptive robustness property is not present in other robust estimators of [12, 13, 18, 21, 28, 32], which involve hard to select tuning parameters. Also none of these estimators is able to identify outliers in the extremes, a property inherent to the adaptive trimmed Hill estimator. An R shiny app implementing the trimmed Hill estimator and the methodology for selection of k_0 is available at [29].

The paper is structured as follows. In Section 2, we study the benchmark Pareto setting. We establish finite-sample optimality and robustness properties of the trimmed Hill estimator. We also introduce a sequential testing method for the automatic selection of k_0. Section 3 deals with the asymptotic properties of the trimmed Hill estimator in the general heavy-tailed regime. The consistency of the sequential testing method is also studied. In Section 4, the finite-sample performance of the trimmed Hill estimator is studied in the context of various heavy tailed models, tail indices, and contamination scenarios. In Sections 4.3, 4.4 and 4.5, we demonstrate the need for adaptive robustness and the advantages of our estimator in comparison with established robust estimators in the literature. In Section 5, we demonstrate the application of the adaptive trimmed Hill methodology to the Condroz data set and French insurance claim settlements data set.

2. Optimal and adaptive trimming: The Pareto regime

In this section, we shall focus on the fundamental Pareto(σ, ξ) model and assume that

$$\mathbb{P}(X > x) = (x/\sigma)^{-1/\xi},\ x \geq \sigma, \tag{2.1}$$

for some $\sigma > 0$ and a tail index $\xi > 0$.

Motivated by the goal to provide a robust estimate of the tail index ξ, we consider trimmed versions of the *classical Hill* estimator in Relation (1.2) and thereby study the class of statistics, $\widehat{\xi}^{\text{trim}}_{k_0,k}(n)$ of Relation (1.3). Proposition 2.1 below finds the optimal weights, $c_{k_0,k}(i)$ for which the estimator, $\widehat{\xi}^{\text{trim}}_{k_0,k}(n)$ is not only unbiased for ξ, but also has the minimum variance. This yields the *trimmed*

Hill estimator of Relation (1.4). Its performance for general heavy-tailed models (see Relation (1.1)) is discussed in Section 3.

2.1. The trimmed Hill estimator

We develop our trimmed Hill estimator as the minimum variance unbiased estimator (MVUE) of ξ among the class of estimators given by Relation (1.3). The class of estimators in Relation (1.3) is linear in terms of the log ratio of order statistics to the k^{th} order statistic. Thus, the following result shows that the trimmed Hill estimator may be viewed as a best linear unbiased estimator (BLUE).

Proposition 2.1. *Suppose $X_1, \cdots, X_n$ are i.i.d.* Pareto(σ, ξ) *random variables, as in Relation (2.1). Then, among the class of linear estimators in Relation (1.3), for $0 \le k_0 < k < n$, the BLUE of ξ is given by*

$$\widehat{\xi}_{k_0,k}(n) = \frac{k_0}{k-k_0} \log\left(\frac{X_{(n-k_0,n)}}{X_{(n-k,n)}}\right) + \underbrace{\frac{1}{k-k_0} \sum_{i=k_0+1}^{k} \log\left(\frac{X_{(n-i+1,n)}}{X_{(n-k,n)}}\right)}_{\widehat{\xi}^{0}_{k_0,k}(n)} \tag{2.2}$$

The proof is given in Section C.2.

Remark 2.2. *The second summand, $\widehat{\xi}^{0}_{k_0,k}(n)$ in Relation (2.2) is nothing but the classic Hill estimator applied to the observations $X_{(n-k_0,n)} \ge \cdots \ge X_{(n-k,n)}$ which denote the top-k order statistics excluding the top-k_0 ones. Note that, $\widehat{\xi}^{0}_{k_0,k}(n)$ which belongs to the class of estimators in Relation (1.3), is not only suboptimal but also biased for the tail index ξ. We shall thus refer to it as the* biased Hill *estimator. The biased Hill estimator has been previously used for robust analysis (see [37]) and inference in truncated Pareto models (see [30], [39]).*

Remark 2.3 (**Classic, Biased and Trimmed Hill Plots**). *The* classic Hill plot *is a plot of the classic Hill estimator, $\widehat{\xi}_k(n)$ as function of k. Likewise, for a fixed k_0, a plot of the trimmed Hill estimator, $\widehat{\xi}_{k_0,k}(n)$ and the biased Hill estimator, $\widehat{\xi}^{0}_{k_0,k}(n)$ as function of k will be referred to as the* trimmed Hill plot *and the* biased Hill plot*, respectively. Since $\widehat{\xi}^{0}_{k_0,k}(n) \le \widehat{\xi}_{k_0,k}(n)$, the biased Hill plot always lies below the trimmed Hill plot. Depending upon the nature of outliers in the extremes, the classic Hill plot can either lie above or below the trimmed Hill plot (see Figures 1 and 11).*

In the rest of the section, we discuss the robustness and finite-sample optimality properties of the trimmed Hill estimator. In this direction, inspired by [12], we define the notion of strict upper breakdown point.

Definition 2.4. *A statistic $\mathcal{T}$ is said to have a strict upper breakdown point β, $0 \le \beta < 1$, if $\mathcal{T} = \mathcal{T}(X_{(n-[n\beta],n)}, \cdots, X_{(1,n)})$ where $X_{(n,n)} \ge \cdots \ge X_{(1,n)}$ are*

the order statistics of the sample, i.e., $\mathcal{T}$ is unaffected by the values of the top $[n\beta]$ order statistics.

In Proposition 2.1, we showed that the trimmed Hill estimator is the BLUE for a large class of estimators with strict upper break down point of k_0/n (see Relation (1.3)). We next prove a stronger result on the finite sample near-optimality of the trimmed Hill estimator. As stated in the next proposition, the trimmed Hill estimator is essentially the minimum variance unbiased estimator (MVUE) among the class of all tail index estimators with a given strict upper break down point.

Theorem 2.5. *Consider the class of statistics given by*

$$\mathcal{U}_{k_0} := \left\{\mathcal{T} = \mathcal{T}(X_{(n-k_0,n)}, \cdots, X_{(1,n)}) : \mathbb{E}(\mathcal{T}) = \xi, \text{ if } X_1, \cdots, X_n \overset{i.i.d.}{\sim} \text{Pareto}(\sigma, \xi)\right\}$$

which are all unbiased estimators of ξ with strict upper breakdown point $\beta = k_0/n$. Then for $\widehat{\xi}_{k_0,n-1}(n)$ as in Relation (2.2), we have

$$\frac{\xi^2}{n-k_0} \leq \inf_{\mathcal{T}\in\mathcal{U}_{k_0}} \text{Var}(\mathcal{T}) \leq \text{Var}(\widehat{\xi}_{k_0,n-1}(n)) = \frac{\xi^2}{n-k_0-1}. \tag{2.3}$$

In particular, $\widehat{\xi}_{k_0,n-1}(n)$ is asymptotically MVUE of ξ among the class of estimators described by $\mathcal{U}_{k_0}$.

The proof is given in Section C.3.

Though the trimmed Hill estimator has nice finite sample properties, it is of limited use in practice unless the value of trimming parameter k_0 is known. In the following section, we will develop a data-driven method for the estimation of k_0.

2.2. Automated selection of the Trimming parameter

In this section, we introduce a methodology for the automated data-driven selection of the trimming parameter k_0. The trimmed Hill estimator with this estimated value of k_0 will be referred to as the *adaptive trimmed Hill* estimator. Its performance as a robust estimator of the tail index ξ is discussed elaborately under Section 4. In addition, the k_0-estimation methodology also provides a tool for the detection of outliers in the extremes of heavy tailed data.

We begin with a result on the joint distribution of the trimmed Hill statistics, which is a starting point towards the estimation of k_0.

Proposition 2.6. *The joint distribution of $\widehat{\xi}_{k_0,k}(n)$ can be expressed as follows:*

$$\left\{\widehat{\xi}_{k_0,k}(n),\ k_0 = 0, \ldots, k-1\right\} \overset{d}{=} \left\{\xi\frac{\Gamma_{k-k_0}}{k-k_0},\ k_0 = 0, \ldots, k-1\right\}, \tag{2.4}$$

where $\Gamma_i = E_1 + \cdots + E_i$ *with* $E_1, E_2, \cdots$ *i.i.d. standard exponential random variables. Consequently, as* $k - k_0 \to \infty$,

$$\sqrt{k-k_0}(\widehat{\xi}_{k_0,k}(n) - \xi) \overset{d}{\Longrightarrow} N(0, \xi^2) \tag{2.5}$$

The proof is given in Section C.2. This result motivates a simple visual device for the selection of k_0.

Diagnostic plot For a fixed value of k, the plot of $\widehat{\xi}_{k_0,k}(n)$ as a function of k_0 will be referred to as a trimmed Hill *diagnostic plot*. The plot also includes additional vertical lines representing $\widehat{\xi}_{k_0,k}(n) \pm \widehat{\sigma}_{k_0,k}(n)$ with $\widehat{\sigma}_{k_0,k}(n) = \widehat{\xi}_{k_0,k}(n)/\sqrt{k-k_0}$.

Note that for observations from ideal Pareto, $\widehat{\sigma}_{k_0,k}(n)$ is indeed the plug in estimate for standard error of $\widehat{\xi}_{k_0,k}(n)$ (see Proposition 2.6). Figure 2, shows diagnostic plots for data simulated from Pareto(1,1) under the case of no outliers (left panel) and $k_0 = 5$ outliers (right panel).

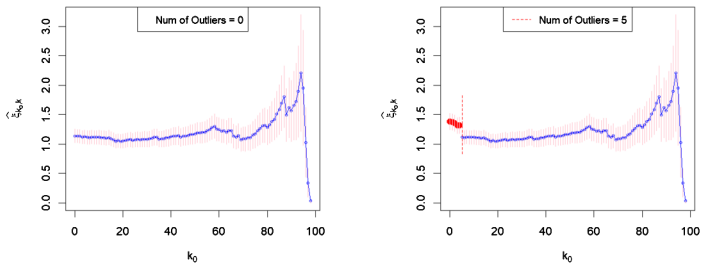

FIG 2. *Diagnostic Plot for Pareto(1,1) with* $n = 100, k = n - 1$. Left*: No outliers.* Right: *5 outliers.*

In the absence of outliers, modulo variability, the diagnostic plot should be constant in k_0 (see left panel in Figure 2). The right panel in Figure 2 corresponds to a case where extreme outliers have been introduced by raising the top-$k_0 = 5$ order statistics to a power greater than 1. This resulted in a visible kink in the diagnostic plot near $k_0 = 5$. Note that, in principle, the presence of outliers could lead to a kink/or change point with an upward or downward trend in the left part of the plot. The diagnostic plot, while useful, requires visual inspection of the data. In practice, an automated procedure is often desirable.

The crux of our methodology for automated selection of k_0 lies in the next result. The idea is to automatically detect a change point in the diagnostic plot by examining it sequentially from right to left. Formally, this will be achieved by a sequential testing algorithm involving the ratio statistics introduced next.

Proposition 2.7. *Suppose all the* X_i*'s are generated from* Pareto(σ, ξ). *Then, the statistics*

$$T_{k_0,k}(n) := \frac{(k-k_0-1)\widehat{\xi}_{k_0+1,k}(n)}{(k-k_0)\widehat{\xi}_{k_0,k}(n)}, \qquad k_0 = 0, 1, \cdots, k-2 \tag{2.6}$$

are independent and follow $\text{Beta}(k-k_0-1,1)$ *distribution for* $k_0 = 0, 1, \cdots, k-2$.

Remark 2.8. *Note that,* $T_{k_0,k}(n)$ *depends only on* $X_{(n-k_0,n)}, \cdots, X_{(n-k,n)}$. *Therefore, the joint distribution of* $T_{k_0,k}(n)$*'s remains the same as long as*

$$(X_{(n-k_0,n)}, \cdots, X_{(n-k,n)}) \stackrel{d}{=} (Y_{(n-k_0,n)}, \cdots, Y_{(n-k,n)})$$

where $Y_{(n,n)} > \cdots > Y_{(1,n)}$ *are the order statistics of* n *i.i.d. observations from* $\text{Pareto}(\sigma, \xi)$. *In other words, Proposition 2.7 holds even in the presence of outliers provided that the outliers are confined only to the top-*k_0 *order statistics. This motivates the sequential testing methodology discussed next.*

Weighted sequential testing By Proposition 2.7, in the Pareto regime, the statistics

$$U_{k_0,k}(n) := 2|(T_{k_0,k}(n))^{k-k_0-1} - 0.5|, \qquad k_0 = 0, 1, \cdots, k-2. \tag{2.7}$$

are i.i.d. $U(0,1)$. This follows from the simple observation that $T_{k_0,k}^{k-k_0-1}(n) \sim U(0,1)$. For simplicity, both in terms of notation and computation, we use the transformation in Relation (2.7) to switch from beta to uniformly distributed random variables.

Assuming that outliers affect only the top-k_0 order statistics, one can identify k_0 as the largest value j for which $U_{j,k}(n)$ fails a test for uniformity. Specifically, we consider a sequential testing procedure, where starting with $j = k-2$, we test the null hypothesis $\mathcal{H}_0(j) : U_{j,k}(n) \sim U(0,1)$ at level α_j. If we fail to reject $\mathcal{H}_0(j)$, we set $j = j-1$ and repeat the process until we either encounter a rejection or $j = 0$. The resulting value of j is our estimate $\widehat{k}_0$. The methodology is formally described in the following algorithm.

Algorithm 1 Weighted Sequential Testing

1: Consider a set of $\alpha_j \in (0,1), j = 0, 1, \cdots, k-2$.
2: Set $j = k-2$.
3: Compute $U_{j,k}(n)$ as in Relation (2.7).
4: If $U_{j,k}(n) < 1-\alpha_j$, set $j = j-1$ else go to step 6.
5: If $j = 0$, go to step 6 else go to step 3.
6: Return $\widehat{k}_0 = j$.

Since α_j varies as a function of j, we refer to Algorithm 1 as the *weighted sequential testing* algorithm. As shown in the following Proposition, the family wise error rate of the algorithm is well calibrated at level $q \in (0,1)$, provided

$$\prod_{j=0}^{k-2}(1-\alpha_j) = 1-q. \tag{2.8}$$

Proposition 2.9. *For i.i.d. observations from Pareto*(σ, ξ)*, let* $\widehat{k}_0$ *be the value from Algorithm 1 with* α_j*'s as in Relation (2.8). Then, under the null hypothesis* $\mathcal{H}_0 : k_0 = 0$, *we have* $\mathbb{P}_{\mathcal{H}_0}(\widehat{k}_0 > 0) = q$.

Proof. We shall instead show, $\mathbb{P}_{\mathcal{H}_0}(\widehat{k}_0 = 0) = 1 - q$. Since the $U_{j,k}(n)$'s are independent $U(0,1)$,

$$\mathbb{P}_{\mathcal{H}_0}(\widehat{k}_0 = 0) = \mathbb{P}_{\mathcal{H}_0}\Big(\bigcap_{j=0}^{k-2}(U_{j,k}(n) < 1 - \alpha_j)\Big) = \prod_{j=0}^{k-2}(1-\alpha_j) = 1 - q, \qquad (2.9)$$

which completes the proof. □

Remark 2.10 (**Choice of** α_j)**.** *For the purposes of this paper, the levels* α_j *in the above algorithm are chosen as follows:*

$$\alpha_j = 1 - (1-q)^{ca^{k-j-1}}, \qquad j = 0, \cdots, k-2 \qquad (2.10)$$

with $a > 1$ *and* $c = 1/\sum_{j=0}^{k-2} a^{k-j-1}$*. This choice of* α_j *satisfies Relation* (2.8)*, which in view of Proposition 2.9, ensures that the algorithm is well calibrated. In addition, this choice puts less weight on large values of* j *and thereby allows for a larger Type I error or fewer rejections for the hypothesis* $\mathcal{H}_0(j) : U_{j,k}(n) \sim U(0,1)$*. This implies that large values of* j *are less likely to be chosen over smaller ones. This guards against encountering spurious values of* $\widehat{k}_0$ *close to* k*, which can lead to highly variable estimates of* $\widehat{\xi}_{k_0,k}(n)$*. Our extensive analysis with a variety of sequential tests indicate that the choice of levels as in Relation* (2.10) *with* $a = 1.2$ *works well in practice.*

Remark 2.11. *Proposition 2.9 shows that in the Pareto case the weighted sequential testing algorithm is well calibrated and attains the exact level Type I error. In the general heavy tailed regime, Theorem 3.11 (below) establishes the asymptotic consistency of the algorithm. In Section 4, we show that the algorithm can identify the true* k_0 *in the ideal Pareto regime as well as the challenging cases of Burr and T distributions (see Section 4.5).*

3. The general heavy tailed regime

In this section, we study the asymptotic properties of $\widehat{\xi}_{k_0,k}$, the trimmed Hill statistic of Relation (2.2), for a general class of heavy-tailed distributions F as in Relation (1.1). Consider the tail quantile function corresponding to F, defined as follows:

$$Q(t) = \inf\{x : F(x) \geq 1 - 1/t\} = F^{-1}(1 - 1/t), \;\; t > 1. \qquad (3.1)$$

Following [4], for F as in Relation (1.1), one can equivalently assume that

$$Q(t) = t^{\xi}\mathcal{L}(t) \qquad (3.2)$$

Remark 3.1. *The relation between the slowly varying functions* ℓ *and* $\mathcal{L}$ *in Relations* (1.1) *and* (3.2) *is well known (see e.g., [4], [11] and [34]). Specifically, one can show that*

$$\mathcal{L}(t^{1/\xi}) \sim \ell^{\xi}(t\mathcal{L}(t^{1/\xi})), \;\; \text{as} \;\; t \to \infty. \qquad (3.3)$$

Thus, $\tilde{\ell}(x) = \ell^{\xi}(x)$ and $\tilde{\mathcal{L}}(x) = \mathcal{L}(x^{1/\xi})$ satisfy $\tilde{\mathcal{L}}(x) \sim \tilde{\ell}(x\tilde{\mathcal{L}}(x))$. This in view of Theorem 1.5.13 of [10] implies that $1/\tilde{\ell}$ is the de Bruijn conjugate *of $\tilde{\mathcal{L}}$ and hence unique up to asymptotic equivalence.*

We start with a conceptually important derivation used in the rest of the section. Using the tail-quantile function, one can express the trimmed Hill statistic under the general heavy-tailed model (1.1) as the sum of a trimmed Hill statistic based on ideal Pareto data plus a remainder term. In view of Relations (3.1) and (3.2), let

$$X_i = Q(Y_i) = Y_i^{\xi}\mathcal{L}(Y_i),\ i = 1, \ldots, n, \tag{3.4}$$

where Y_i's i.i.d Pareto$(1, 1)$. Then $X_i,\ i = 1, \ldots, n$, represent an i.i.d. sample from F.

Therefore, using $X_{(n-i,n)} = Q(Y_{(n-i,n)})$ in Relation (2.2), we obtain

$$\begin{aligned}\widehat{\xi}_{k_0,k}(n) &= \underbrace{\frac{k_0}{k-k_0}\log\frac{Y^{\xi}_{(n-k_0,n)}}{Y^{\xi}_{(n-k,n)}} + \frac{1}{k-k_0}\sum_{i=k_0}^{k-1}\log\frac{Y^{\xi}_{(n-i,n)}}{Y^{\xi}_{(n-k,n)}}}_{\widehat{\xi}^{*}_{k_0,k}(n)} \\ &+ \underbrace{\frac{k_0}{k-k_0}\log\frac{\mathcal{L}(Y_{(n-k_0,n)})}{\mathcal{L}(Y_{(n-k,n)})} + \frac{1}{k-k_0}\sum_{i=k_0}^{k-1}\log\frac{\mathcal{L}(Y_{(n-i,n)})}{\mathcal{L}(Y_{(n-k,n)})}}_{R_{k_0,k}(n)}\end{aligned} \tag{3.5}$$

where $Y_{(i,n)}$'s are the order statistics for the Y_i's. Since Y_i^{ξ}'s follow Pareto$(1, \xi)$, the statistic $\widehat{\xi}^{*}_{k_0,k}(n)$ in Relation (3.5) is simply the trimmed Hill estimator for ideal Pareto data and $R_{k_0,k}(n)$ is a remainder term that encodes the effect of the slowly varying function $\mathcal{L}$.

The nature of the function $\mathcal{L}$ determines the rate at which the remainder term $R_{k_0,k}(n)$ converges to 0 in probability. We establish the minimax rate optimality of the trimmed Hill estimator under the Hall class of assumptions on the function $\mathcal{L}$ (see Section 3.1). To establish the asymptotic normality of the trimmed Hill estimator, we use second order regular variation conditions on the function $\mathcal{L}$ (see Section 3.2). Under the same set of conditions, the asymptotic consistency of the weighted sequential testing algorithm of Section 2.2 is also established in Section 3.3.

3.1. Minimax rate optimality of the trimmed Hill estimator

Here, we study the rate-optimality of the trimmed Hill estimator for the class of distributions in $\mathcal{D} := \mathcal{D}_{\xi}(B, \rho)$, where Relation (3.2) holds with tail index $\xi > 0$ and $\mathcal{L}$ of the form:

$$\mathcal{L}(x) = 1 + r(x), \qquad \text{with} \qquad |r(x)| \le Bx^{-\rho},\ (x > 0) \tag{3.6}$$

for constants $B > 0$ and $\rho > 0$ (see also Relation (2.7) in [11]). This is known as the *Hall class* of distributions.

In [24], Hall and Welsh showed that no estimator can be uniformly consistent over the class of distributions in $\mathcal{D}$ at a rate faster than or equal to $n^{\rho/(2\rho+1)}$. Theorem 1 of [24] adapted to our setting and notation is as follows:

Theorem 3.2 (optimal rate)**.** *Let $\widehat{\xi}_n$ be any estimator of ξ based on an independent sample from a distribution $F \in \mathcal{D}_\xi(B,\rho)$. If we have*

$$\liminf_{n\to\infty} \inf_{F\in\mathcal{D}_\xi(\mathcal{B},\rho)} \mathbb{P}_F(|\widehat{\xi}_n - \xi| \le a(n)) = 1 \tag{3.7}$$

then $\liminf_{n\to\infty} n^{\rho/(2\rho+1)}a(n) = \infty$. Here by $\mathbb{P}_F$, we understand that $\widehat{\xi}_n$ was based on independent realizations from F.

In Theorem 3 of [24], it is shown that for the case of no outliers, the classic Hill estimator, $\widehat{\xi}_k(n)$ with $k = k(n) \sim n^{2\rho/(1+2\rho)}$ is a uniformly consistent estimator of ξ at a rate greater than or equal to any other uniformly consistent estimator. In other words, the classic Hill estimator is minimax rate optimal in view of Theorem 3.2 wherein $\widehat{\xi}_n = \widehat{\xi}_{k(n)}$ satisfies Relation (3.7) for every $a(n)$ with $a(n)n^{\rho/(2\rho+1)} \to \infty$.

Note that, Theorem 3.2 also applies to the trimmed Hill estimator. We next show that in the presence of outliers, the trimmed Hill estimator with $k_0 = k_0(n) = o(k)$ and $k = k(n) \sim n^{2\rho/(1+2\rho)}$ is minimax rate optimal with the same rate as that of the classic Hill. In addition, the minimax rate optimality holds uniformly over all $k_0 \in [0, h(k)]$ for $h(k) = o(n^{2\rho/(1+2\rho)})$.

Theorem 3.3 (uniform consistency)**.** *Suppose that $k = k(n) \propto n^{2\rho/(2\rho+1)}$ and $h(k) = o(k)$, as $n \to \infty$.*

Then, for every sequence $a(n) \downarrow 0$, such that $a(n)\sqrt{k(n)} \to \infty$, we have

$$\liminf_{n\to\infty} \inf_{F\in\mathcal{D}_\xi(B,\rho)} \mathbb{P}_F\left(\max_{0\le k_0<h(k)} |\widehat{\xi}_{k_0,k}(n) - \xi| \le a(n)\right) = 1. \tag{3.8}$$

The proof of this result is given in Section C.3.1. Observe that $\sqrt{k(n)} \propto n^{\rho/(1+2\rho)}$ is the optimal rate in Theorem 3.2. Therefore, Theorem 3.3 implies that $\widehat{\xi}_{k_0,k}(n)$ is minimax rate-optimal in the sense of Hall and Welsh [24]. Also, note that the trimmed Hill estimator $\widehat{\xi}_{k_0,k}(n)$ is *uniformly consistent* with respect to both the family of possible distributions $\mathcal{D}$ as well as the trimming parameter k_0, provided $k_0 \in [0, h(k)]$ for $h(k) = o(k)$.

Remark 3.4. *The above appealing result shows that trimming does not sacrifice the rate of estimation of ξ so long as $k_0 = o(n^{2\rho/(2\rho+1)})$, $n \to \infty$. In the regime where the rate of contamination k_0 exceeds $n^{2\rho/(2\rho+1)}$, to achieve robustness and asymptotic consistency, one would have to choose $k(n) \gg n^{2\rho/(2\rho+1)}$, which naturally leads to rate-suboptimal estimators. In this case, similar uniform consistency for the trimmed Hill estimators can be established along the lines of Theorem 3.3.*

3.2. Asymptotic normality of the trimmed Hill estimator

Here, we shall establish the asymptotic normality of $\widehat{\xi}_{k_0,k}(n)$ under the general semi-parametric regime (Relation (1.1) or equivalently Relation (3.2)). In Proposition 2.6, we already established the asymptotic normality of the trimmed Hill estimator in the Pareto regime. Recalling Relation (3.5), we observe that $\widehat{\xi}_{k_0,k}(n)$ differs from a tail index estimator based on Pareto data only by a remainder term $R_{k_0,k}(n)$. Thus, proving the asymptotic normality of $\widehat{\xi}_{k_0,k}(n)$ amounts to controlling the asymptotic behavior of the remainder term.

Indeed, we begin with a much stronger result which establishes the convergence rate of $R_{k_0,k}(n)$ *uniformly* for all $k_0 \in [0, h(k)]$ where $h(k) \in o(k)$. To this end, following [4], we adopt the following second order condition on the function $\mathcal{L}$:

$$\sup_{t \geq t_\varepsilon} \left| \log \frac{\mathcal{L}(tx)}{\mathcal{L}(t)} - cg(t) \int_1^x \nu^{-\rho-1} d\nu \right| \leq \begin{cases} \varepsilon g(t) & \text{if } \rho > 0 \\ \varepsilon g(t) x^\varepsilon & \text{if } \rho = 0. \end{cases} \tag{3.9}$$

for all $\varepsilon > 0$ and some t_ε dependent on ε and $g : (0, \infty) \to (0, \infty)$ is a $-\rho$ varying function with $\rho \geq 0$ (see Lemma A.2 in [4] for more details).

Theorem 3.5. *Suppose the X_i's are independent realizations with tail quantile function Q as in Relation (3.2) with $\mathcal{L}$ as in Relation (3.9). If, for some $\delta > 0$ and constant $A > 0$,*

$$k^\delta g(n/k) \to A \text{ for } k/n \to 0 \text{ as } k, n \to \infty, \tag{3.10}$$

then for $R_{k_0,k}(n) = \widehat{\xi}_{k_0,k}(n) - \widehat{\xi}^*_{k_0,k}(n)$ as in Relation (3.5) and $h(k) = o(k)$, we have

$$k^\delta \max_{0 \leq k_0 < h(k)} \left| R_{k_0,k}(n) - \frac{cAk^{-\delta}}{1+\rho} \right| \xrightarrow{\mathbb{P}} 0 \tag{3.11}$$

The proof is given in Section C.3.2.

Remark 3.6. *Simulation results of Section A show that the rate δ in Relation (3.11) is indeed optimal for the Hall class of distributions (Relation (3.6)) and cannot be improved further (see Table 8). The exact proof is however beyond the scope of this paper.*

The asymptotic normality of $\widehat{\xi}_{k_0,k}(n)$ is a direct consequence of Theorem 3.5 with $\delta = 1/2$ and Relation (2.5). This is formalized in the following corollary.

Corollary 3.7. *If $k_0 = o(k)$ and $\sqrt{k} g(n/k) \to A \in [0, \infty)$,*

$$\sqrt{k}(\widehat{\xi}_{k_0,k}(n) - \xi) \overset{d}{\Longrightarrow} N\left(\frac{cA}{1+\rho}, \xi^2 \right), \quad \text{as } n \to \infty. \tag{3.12}$$

Proof. By adding and subtracting the estimator $\widehat{\xi}^*_{k_0,k}(n)$ defined in Relation (3.5), we have

$$\begin{aligned}\sqrt{k}(\widehat{\xi}_{k_0,k}(n)-\xi) &= \sqrt{k}(\widehat{\xi}_{k_0,k}(n)-\widehat{\xi}^*_{k_0,k}(n))+\sqrt{k}(\widehat{\xi}^*_{k_0,k}(n)-\xi)\\ &= \sqrt{k}R_{k_0,k}(n)+\sqrt{k}(\widehat{\xi}^*_{k_0,k}(n)-\xi). \qquad (3.13)\end{aligned}$$

For the first term in Relation (3.13), we have $\sqrt{k}R_{k_0,k} \xrightarrow{\mathbb{P}} cA/(1+\rho)$. This follows from Relation (3.11) with $\delta = 1/2$. By Relation (2.5), the second term in Relation (3.13) satisfies $\sqrt{k}(\widehat{\xi}^*_{k_0,k}-\xi) \xrightarrow{d} N(0,\xi^2)$, as $k\to\infty$, and hence (3.12) follows. □

Remark 3.8. *Consider the asymptotic normality result of Corollary 3.7 for the Hall class of distributions in Relation (3.6). In this case, we have $g(x)\propto x^{-\rho}$ and the convergence $\sqrt{k}g(n/k)\to A>0$ implies that $k=k(n)\propto n^{2\rho/(2\rho+1)}$, as $n\to\infty$. This is the optimal rate, which as we know from Theorem 3.3, cannot be achieved by an asymptotically unbiased estimator of ξ. Indeed, the limit distribution in Relation (3.12) involves the bias term $cA/(\rho+1)$. To eliminate the bias term, one can pick $k=o(n^{2\rho/(2\rho+1)})$, which in this case implies that $\sqrt{k}g(n/k)\to A\equiv 0$. That is, asymptotically unbiased estimators can be obtained but one needs to sacrifice the optimal rate.*

3.3. Asymptotic behavior of the weighted sequential testing

In this section, we establish the asymptotic consistency of the weighted sequential testing algorithm (Algorithm 1) under the same set of second order regular variation conditions on the function $\mathcal{L}$ as in Section 3.2. We begin with a convergence result on the ratio statistics of Relation (2.6).

Theorem 3.9. *Assume that the conditions of Theorem 3.5 hold and Relation (3.10) holds for some $\delta>0$, then*

$$k^{\delta}\max_{0\le k_0<h(k)}\left|T_{k_0,k}(n)-T^*_{k_0,k}(n)\right| \xrightarrow{\mathbb{P}} 0, \qquad (3.14)$$

where $T_{k_0,k}(n)$ and $T^*_{k_0,k}(n)$ are based on $\widehat{\xi}_{k_0,k}(n)$ and $\widehat{\xi}^*_{k_0,k}(n)$, respectively as in Relation (2.6).

The proof is described in Section C.3.3.

Remark 3.10. *The question of whether the rate δ in Relation (3.14) can be improved or not is unresolved. Simulation results in Section A (see Table 8) show that the achievable optimal rate is indeed greater than δ. The exact proof of this is however more involved requiring additional assumptions on the slowly varying function $\mathcal{L}$ of Relation (3.2).*

We next establish that the weighted sequential testing algorithm is well calibrated and attains the significance level q even for the general class of heavy tailed models in Relation (1.1).

Theorem 3.11. *Suppose Relation (3.14) in Theorem 3.9 holds with $\delta = \delta^*$, then*

1. *Based on $T_{k_0,k}(n)$ and $T^*_{k_0,k}(n)$, suppose $U_{k_0,k}(n)$ and $U^*_{k_0,k}(n)$ are defined as in Relation (2.7), then*

$$k^{(\delta^*-1)} \max_{0\le k_0<h(k)} \left| U_{k_0,k}(n) - U^*_{k_0,k}(n) \right| \xrightarrow{\mathbb{P}} 0, \tag{3.15}$$

2. *Suppose in Step 2 of Algorithm 1, j starts from $h(k)$ instead of $k-2$, then under $\mathcal{H}_0 : k_0 = 0$*

$$\mathbb{P}_{\mathcal{H}_0}[\widehat{k}_0 > 0] \longrightarrow q. \tag{3.16}$$

as long as $\delta^* \ge 2$.

The proof is given in the Section C.3.3.

Remark 3.12. *If Relation (3.14) holds at a rate $\delta = \delta^*$, there is very little scope of improvement in rates for Relations (3.15) and (3.16). Extensive simulation results of Section A demonstrate that the rates of Theorem 3.11 as derived from the rate in Theorem 3.9 are indeed optimal (see Table 8 and Figure 12) at least for the Hall class of Relation (3.6).*

Remark 3.13. *For the Hall class of distributions, $g(x) \sim x^{-\rho}$, thus Relation (3.10) holds whenever*

$$k = k(n) \propto n^{\rho/(\rho+\delta)}. \tag{3.17}$$

For this choice of k, by Theorem 3.9, the $T_{k_0,k}(n)$'s converge at least at the rate δ. This implies that $U_{k_0,k}(n)$'s and the Type I error of Algorithm 1 converge at least for $\delta \ge 1$ and $\delta \ge 2$, respectively[1]. By Remark 3.8, the minimax rate optimality and the asymptotic normality of the trimmed Hill estimator $\widehat{\xi}_{k_0,k}$ with $\delta = 1/2$ in Relation (3.17). Choices of $\delta \ge 1$ and $\delta \ge 2$ produce suboptimal choices of k in terms of rate. If ρ is large, the difference between these suboptimal values and the optimal value $n^{\rho/(\rho+1/2)}$ is negligible. For small values of ρ, the difference is greater and the consistency of Algorithm 1 may be compromised. However, in Section 4.5, we show that even at smaller values of ρ, we do a reasonably good job in terms of determining the true k_0.

4. Performance of the adaptive trimmed Hill estimator

4.1. Simulation set up

In this section, we study the finite sample performance of the adaptive trimmed Hill estimator, $\widehat{\xi}_{\widehat{k}_0,k}(n)$, which is the trimmed Hill statistic in Relation (2.2) with $k_0 = \widehat{k}_0$ (see also the R shiny app at [29]). Here, the value of the trimming parameter $\widehat{k}_0$ is obtained from the weighted sequential testing algorithm, Algo-

[1] We believe that the conditions $\delta \ge 1$ and $\delta \ge 2$ can be made less stringent since there is scope of improvement in the rate δ for $T_{k_0,k}$ in Relation (3.14).

rithm 1. We also evaluate the accuracy of the algorithm as an estimator of the number of outliers k_0. The parameters for the algorithm, a and q are set at 1.2 and 0.05 respectively.

Measures of performance The performance of an estimator $\widehat{\xi}$ of ξ is evaluated in terms of its root mean squared error ($\sqrt{MSE}$), where

$$\mathrm{MSE}(\widehat{\xi}) = \mathbb{E}(\widehat{\xi} - \xi)^2. \tag{4.1}$$

Using criterion (4.1), we evaluate the performance of the adaptive trimmed Hill estimator and several other competing estimators of the tail index ξ. The computation of the $\sqrt{MSE}$ is based on 2500 independent Monte Carlo simulations.

Data generating models We generate n i.i.d. observations from one of the following heavy-tailed distributions:

$$\begin{aligned}
\mathrm{Pareto}(\sigma,\xi) \quad &: \quad 1 - F(x) = \sigma^{1/\xi} x^{-1/\xi}; \;\; x > \sigma, \sigma > 0, \xi > 0, \rho = \infty; \\
\mathrm{Burr}(\eta,\lambda,\xi) \quad &: \quad 1 - F(x) = 1 - \left(\frac{\eta}{\eta + x^{-1/\xi}}\right)^{-\lambda} \\
&; \quad x > 0, \eta > 0, \lambda > 0, \xi > 0, \rho = 1 \\
|\mathrm{T}|(\xi) \quad &: \quad 1 - F(x) = \int_x^\infty \frac{2\Gamma(\frac{1/\xi+1}{2})}{\sqrt{n\pi}\Gamma(\frac{1}{2\xi})} \left(1 + w^2\xi\right)^{-\frac{1+\xi}{2\xi}} dw \\
&; \quad x > 0, \xi > 0, \rho = 2\xi
\end{aligned} \tag{4.2}$$

Sections 4.3 and 4.4 deal with the performance of the weighted sequential testing algorithm and the adaptive trimmed Hill estimator for Pareto observations. Section 4.5 delve deeper into the performance under challenging cases of non Pareto scenarios like the |T| and the Burr distributions.

Choice of k In [25], Hall and Welsh proved that the asymptotic mean squared error of the classic Hill estimator $\widehat{\xi}_k(n)$ is minimal for

$$k_n^{\mathrm{opt}} \sim \left(\frac{C^{2\rho}(\rho+1)^2}{2D^2\rho^3}\right)^{1/(2\rho+1)} n^{2\rho/(2\rho+1)} \tag{4.3}$$

for the Hall class of Relation (3.6). Thus, k_n^{opt} provides an optimal choice of k for computing the classic Hill estimator for data arising from the Hall class of distributions. For the Hall class, in Theorem 3.3, we showed that the trimmed Hill estimator is also optimal at the same rate as the classic Hill estimator as long as the number of outliers, $k_0 = o(k)$. Since for Pareto $\rho = \infty$, the optimal k in view of Relation (4.3) is $n - 1$. Sections 4.3 and 4.4 which deal only with Pareto examples use this value of k. For Sections 4.2 and 4.5 which deal with non-Pareto examples as well, k is chosen approximately around the optimal value as in Relation (4.3).

In Section 4.2, we demonstrate the performance of the adaptive trimmed Hill estimators in the regime of no outliers. In this scenario, the classic Hill estimator (recall Relation (1.2)) is an asymptotically optimal estimator of ξ (see [26]) and is therefore used as the comparative baseline.

Outlier scenarios In Sections 4.3, 4.4 and 4.5, we demonstrate the performance of the adaptive trimmed Hill estimator in the presence of outliers. We next discuss the mechanism of outlier injection to introduce outliers in the extreme observations of the data.

1. *Exponentiated Outliers:* The top-k_0 order statistics are perturbed as follows:

$$X_{(n-i+1,n)} := X_{(n-k_0,n)} + (X_{(n-i+1,n)} - X_{(n-k_0,n)}))^L, \qquad i = 1, \cdots, k_0, \tag{4.4}$$

2. *Scaled Outliers:* The top-k_0 order statistics are perturbed as follows:

$$X_{(n-i+1,n)} := X_{(n-k_0,n)} + C(X_{(n-i+1,n)} - X_{(n-k_0,n)})), \qquad i = 1, \cdots, k_0, \tag{4.5}$$

3. *Mixed Outliers:* For $\mathcal{S} = \{s : X_s > \tau\}$, set

$$\{X_s\}_{s\in\mathcal{S}} = M\tau, \qquad M > 1 \tag{4.6}$$

i.e., observations above a given threshold τ are perturbed.

Firstly note that the L, C and M are constants whose values can be changed to control the intensity of the injected outliers. Secondly, all the above three nature of outliers preserve the order of the bottom-$(n - k_0)$ order statistics. Unlike the case of mixed outliers, the exponentiated and scaled outliers preserve the order of the top-k_0 order statistics as well. The case of mixed outliers is a challenging one because the trimming parameter k_0, though controlled by τ, is random and not well defined. In contrast, k_0 is fixed and well defined for exponentiated and scaled outliers. Thus, for exponentiated and scaled outliers, we demonstrate the efficiency of Algorithm 1 in determining k_0.

Competing robust estimators In the presence of outliers, the adaptive trimmed Hill estimator is indeed a robust estimator of the tail index ξ. Thus, for a comparative baseline we use two other robust estimators of the tail index in Sections 4.3, 4.4 and 4.5. These are the optimal B-robust estimator (OBRE) of [38] and the generalized median estimator (GME) of [12]. For a parametric model, the asymptotic relative efficiency (ARE) of an estimator is defined as the ratio of its asymptotic variance to that of the maximum likelihood estimator (MLE) expressed as a percentage. The tuning parameters of OBRE and GME are chosen such that the ARE can be controlled at a given level. Two values of ARE levels viz 78% and 94% are to allow for varying degrees of robustness. Section B explains the form of these two estimators in addition to the connection between the ARE levels and choice of tuning parameters.

The constant $\widetilde{c}$ which serves as a bound on the influence function (IF) controls the degree of robustness for optimal B-robust estimator (see Relations (B.2) and (B.3) in Section B.1). Indeed, the values $\widetilde{c} = 1.63$ and $\widetilde{c} = 2.73$ result in 78% and 94% asymptotic relative efficiency (ARE), respectively, for the optimal B-robust estimator. Similarly, the parameter κ which controls for the subset size in defining the generalized median statistic also controls for its degree of robustness

(see Relations (B.6) and (B.7) in Section B.2). Indeed, the values $\kappa = 2$ and $\kappa = 5$ produce ARE values 78% and 94%, respectively, for the generalized median estimator. Other robust estimators of the tail index like the probability integral transform statistic estimator of [20] and the partial density component estimator of [36] were also considered but their results are similar and have been omitted for brevity.

4.2. Case of no outliers

For the three distribution models in Relation (4.2), we report the performance of the adaptive trimmed Hill estimator (ADAP) under the regime of no outliers. The classic Hill estimator (HILL) is used as the comparative baseline. For a sample of $n = 1000$ data points, Figure 3 gives the $\sqrt{MSE}$ values for the ADAP and the HILL as a function of k for the distributions Pareto(1,ξ), Burr(1,0.5,ξ), |T|(ξ) with tail index $\xi = 2$. The value of k in Relation (4.3) leads to the smallest $\sqrt{MSE}$ for the HILL. This explains the occurrence of a minima in the plot of $\sqrt{MSE}$ values.

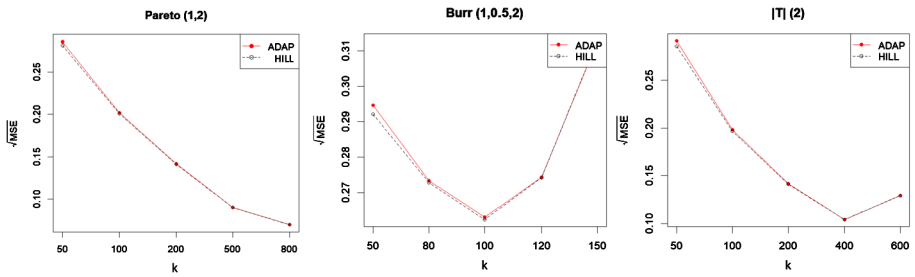

FIG 3. $\sqrt{MSE}$ *of the ADAP for* $\xi = 2$ *and* $k_0 = 0$.

TABLE 1
Type I error of the weighted sequential testing algorithm for $k_0 = 0$

Pareto(1,2)	k=50	k=100	k=200	k=500	k=800
	0.0500	0.0472	0.0496	0.0448	0.0556
Burr(1,0.5,2)	k=50	k=80	k=100	k=150	k=200
	0.0476	0.0496	0.0416	0.0408	0.0392
|T|(2)	k=50	k=100	k=200	k=400	k=600
	0.0484	0.0556	0.0472	0.0520	0.0464

We observe that for a wide range of k, the ADAP is virtually indistinguishable from the HILL irrespective of the distribution under study. This indicates that the weighted sequential testing algorithm can precisely determine $k_0 = 0$ for the same wide range of k-values as in Figure 3. This encouraging finite sample performance complements the theoretically established consistency of the algorithm in Theorem 3.11. Indeed, Table 1 shows that the algorithm attains the nominal significance level of $q = \mathbb{P}(\widehat{k}_0 > 0) = 0.05$.

4.3. Adaptive robustness

In this section, we study how the presence of outliers in the data influences the performance of the adaptive trimmed Hill estimator (ADAP) and the weighted sequential testing algorithm. For clarity and simplicity, the data in this section are generated from Pareto as in Relation (4.2) with $\sigma = 1, \xi = 2$ for varying sample sizes $n = 100, 300, 500$.

The value of k is fixed at $n - 1$ which is indeed the optimal k for the Pareto regime (see Relation (4.3)). Section 4.5 illustrates the adaptive robustness phenomenon as explained in this section in the context of general heavy tailed models as in Relation (1.1). Outliers are injected by Relations (4.4), (4.5) and (4.6), with $L = 3$, $C = 200$ and $M = 100$, respectively. Varying values of the parameter k_0 and τ are chosen to control for the number of outliers in the data.

Figures 4, 5 and 6 produce a plot of the $\sqrt{MSE}$ for the ADAP for outlier generating mechanisms in Relations (4.4), (4.5) and (4.6), respectively.

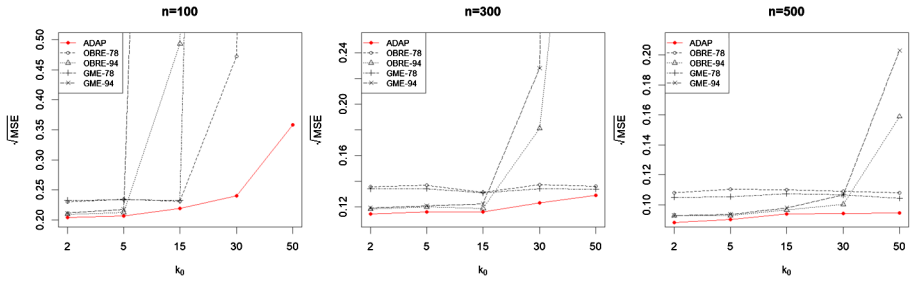

FIG 4. $\sqrt{MSE}$ *of ADAP for Pareto(1,2) with exponentiated outliers:* $L = 3$*, varying* k_0.

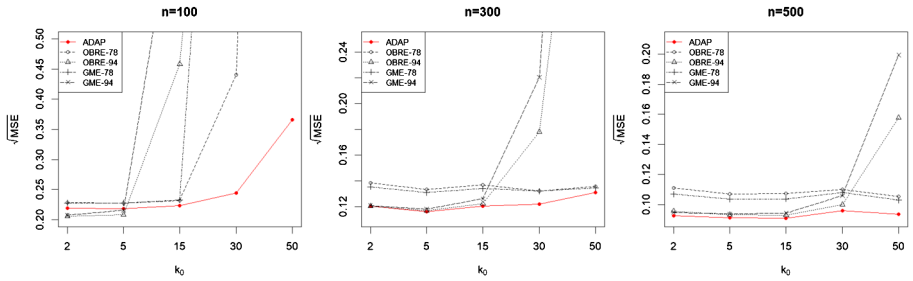

FIG 5. $\sqrt{MSE}$ *of ADAP for Pareto(1,2) with scaled outliers:* $C = 200$*, varying* k_0.

For comparison, the performance of the optimal B-robust estimator (OBRE) and the generalized median estimator (GME) at 78% and 94% ARE levels have also been included. The figures clearly show that the ADAP is uniformly the best estimator in terms $\sqrt{MSE}$. The figures also show an intriguing adaptive robustness property of our estimator. Namely, its $\sqrt{MSE}$ is nearly flat and grows

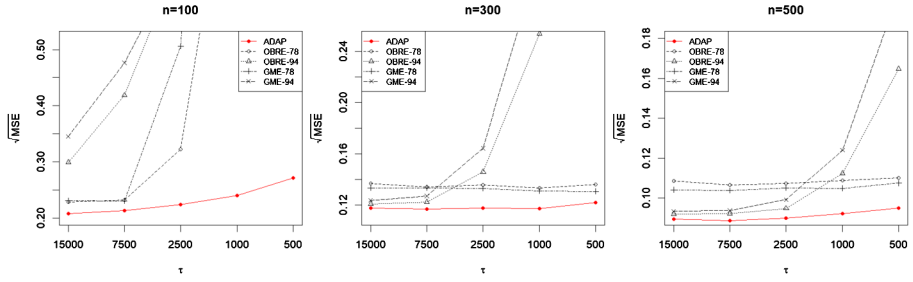

FIG 6. $\sqrt{MSE}$ *of ADAP for Pareto(1,2) with mixed outliers:* $M = 100$*, varying* τ*.*

slowly with increase in the degree of contamination (parametrized by either the number of outliers k_0 in Figures 4 and 5 or the threshold τ in Figure 6). On the other hand, the competing estimators break down completely with increase in the degree of contamination. This can be explained as: the competing estimators must be calibrated to a predefined level of robustness by setting their ARE level in advance. To the best of our knowledge, none of the existing works in the literature provide a data-driven method for selecting this optimal ARE value. In contrast, the trimming parameter k_0 involved in the ADAP is estimated from the data itself which allows it to adapt itself to unknown degrees of contamination in the data.

Figures 4 and 5 show that whenever the target ARE value is greater than $(1-k_0/n)\times 100\%$, the performance of the ADAP is much superior to that of the competing estimators. For example, the OBRE-94 and the GME-94 breakdown completely when $1 - k_0/n \leq 0.9$ ($n = 100, k_0 \geq 15$ and $n = 300, k_0 \geq 30$). Similarly, the performance of the OBRE-78 and the GME-78 is drastically poor where $1 - k_0/n \leq 0.7$ ($n = 100, k_0 \geq 30$). If the target ARE of two estimators is less than $(1 - k_0/n) \times 100\%$, then the estimator with greater ARE has higher efficiency. This explains why the performance of the OBRE-78 and the GME-78 is quite poor in comparison to that of the OBRE-94 and the GME-94 when $1 - k_0/n \geq 0.95$ ($n = 100, k_0 \leq 5$ and $n = 300, k_0 \leq 15$).

By automatically estimating the number of outliers, ADAP not only produces an estimator of ξ robust to varying levels of data contamination but also provides a methodology for outlier detection in the extremes of heavy tailed models. Indeed, Tables 2 and 3 which produce the mean and standard errors of $\widehat{k}_0$ for outliers injected by mechanisms (4.4) and (4.5), show that for all values of n, the weighted sequential testing algorithm picks up the true number of outliers k_0 for almost all values k_0 (exception is $k_0 = 2$ for scaled outliers).

TABLE 2
$\mathbb{E}(\widehat{k}_0) \pm$ Standard Error$(\widehat{k}_0)$ *for Pareto(1,2) with exponentiated outliers,* $L = 3$

n	$k_0 = 2$	$k_0 = 5$	$k_0 = 15$	$k_0 = 30$	$k_0 = 50$
100	2.19 ± 1.42	5.10 ± 1.04	14.99 ± 0.51	29.84 ± 0.41	49.47 ± 0.78
300	2.23 ± 1.61	5.08 ± 0.95	14.98 ± 0.44	29.85 ± 0.44	49.55 ± 0.70
500	2.17 ± 1.19	5.20 ± 3.95	14.98 ± 0.49	29.85 ± 0.39	49.55 ± 0.70

TABLE 3
$\mathbb{E}(\widehat{k}_0) \pm$ Standard Error$(\widehat{k}_0)$ *for Pareto(1,2) with scaled outliers,* $C = 200$

n	$k_0 = 2$	$k_0 = 5$	$k_0 = 15$	$k_0 = 30$	$k_0 = 50$
100	1.10 ± 2.09	4.66 ± 1.87	14.91 ± 0.90	29.89 ± 0.70	49.68 ± 3.01
300	1.06 ± 1.85	4.68 ± 1.75	14.94 ± 1.02	29.91 ± 0.84	49.88 ± 0.39
500	1.09 ± 1.96	4.69 ± 1.83	14.91 ± 0.82	29.97 ± 2.81	49.89 ± 0.37

4.4. Impact of outlier severity and tail index

In this section, we study the influence of the magnitude of outliers and tail index on the performance of the adaptive trimmed Hill estimator (ADAP) for Pareto observations with sample size $n = 500$. The conclusions were similar for other heavy tailed models explored.

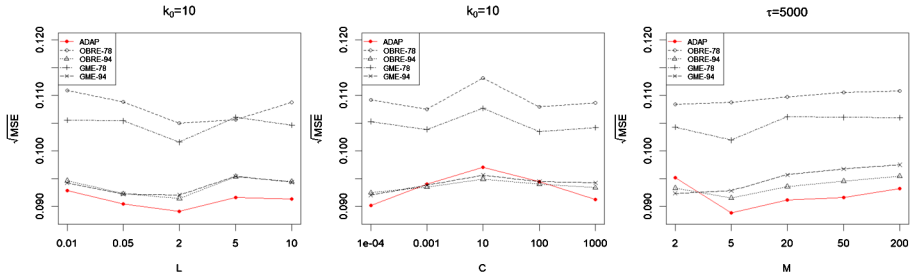

FIG 7. $\sqrt{MSE}$ *of ADAP for Pareto(1,2).* Left: *Exponentiated outliers with varying* L. Middle: *Scaled outliers with varying* C. Right: *Mixed outliers with varying* M.

We begin with the impact of outlier severity on the performance of the ADAP. For outlier generating mechanisms in Section 4.1, the outlier severity is controlled by the parameters L, C and M. The data generating model is Pareto as in Relation (4.2) with $\sigma = 1$ and $\xi = 2$. Figure 7 produces a plot of the $\sqrt{MSE}$ for ADAP for outlier generating mechanisms in Relations (4.4), (4.5) and (4.6) with $k_0 = 10$, $\tau = 5000$ and varying L, C and M. For comparison, $\sqrt{MSE}$ values for the optimal B-robust estimator (OBRE) and the generalized median (GME) at 78% and 94% ARE levels have also been included. The ADAP performs better than both the OBRE and the GME for almost all values of L, C and M no matter what their ARE levels is. The only exception is $C = 10$ for the case scaled outliers (see Relation (4.5)). Though more robust, the estimators the OBRE-78 and the GME-78 perform poorly at lower levels of contamination in the data. This explains their overall inferior behavior in Figure 7 right panel where the degree of contamination is only 2% ($n = 500, k_0 = 10$).

The superiority of the ADAP grows with increase in the severity of the outliers. For exponentiated and scaled outliers, the increase in severity is manifested through an increase in L, C for $L, C > 1$ and decrease in L, C for $L, C < 1$. For mixed outliers, the increase in severity occurs with increase in the value of M. With an increase in severity of outliers, the weighted sequential testing algo-

TABLE 4
$\mathbb{E}(\widehat{k}_0) \pm$ Standard Error$(\widehat{k}_0)$ *for Pareto (1,2) with* $k_0 = 10$ *outliers*

Exp outliers					
L	0.01	0.05	2	5	10
$\mathbb{E}(\widehat{k}_0) \pm \text{SE}(\widehat{k}_0)$	9.74 ± 1.06	9.59 ± 1.11	9.91 ± 0.51	10.05 ± 0.87	10.05 ± 0.71
Scl outliers					
C	0.0001	0.001	10	100	1000
$\mathbb{E}(\widehat{k}_0) \pm \text{SE}(\widehat{k}_0)$	10.05 ± 0.70	8.83 ± 2.17	3.86 ± 4.73	9.57 ± 1.83	10.03 ± 0.61

TABLE 5
$\mathbb{E}(\widehat{k}_0) \pm$ Standard Error$(\widehat{k}_0)$ *for Pareto (1,*ξ*) with* $k_0 = 10$ *outliers for* $L = 3$ *and* $C = 200$

	$\xi = 0.25$	$\xi = 0.5$	$\xi = 1$	$\xi = 1.5$	$\xi = 2.5$
Exp outliers	5.68 ± 4.23	7.33 ± 2.32	9.57 ± 1.03	9.95 ± 0.73	10.03 ± 0.51
Scl outliers	10.00 ± 0.59	10.01 ± 1.18	9.98 ± 0.72	9.99 ± 1.05	9.79 ± 1.47

rithm can correctly detect the true number of outliers k_0 (see Table 4) and hence the greater efficiency of ADAP.

We next study the impact of the tail index ξ on the performance of ADAP. The data generating model is Pareto as in Relation (4.2) with $\sigma = 1$ and varying values of ξ. Outliers are injected according to Relations (4.4), (4.5) and (4.6) with $k_0 = 10$, $\tau = 5000$, $L = 3$, $C = 200$ and M. Figure 8 produces a plot of the $\sqrt{MSE}$ values for the ADAP along with those of the OBRE and the GME at 78% and 94% ARE levels. The performance of the ADAP is superior to that of the remaining estimators. For exponentiated and mixed outliers, the improvement is even more prominent at larger values of ξ. This is because for the same values of L and M, the severity of outliers is greater for heavier tails ($\xi = 2.5$) than lighter ones ($\xi = 0.5$). In contrast, for scaled outliers, the improvement is more prominent at smaller ξ values. This is because for the same value of C, the severity of outliers is greater for lighter tails than heavier ones. This is in consensus with the findings of Table 5 where the accuracy of the weighted sequential testing algorithm, Algorithm 1 in correctly estimating the true number of outliers improves with increase in ξ for exponentiated and mixed outliers and decrease in ξ for scaled outliers.

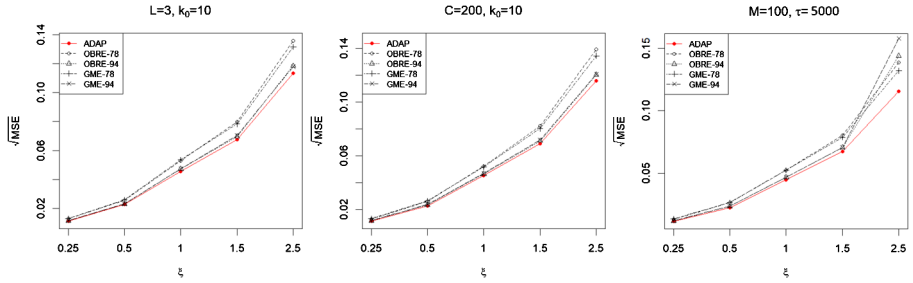

FIG 8. $\sqrt{MSE}$ *of ADAP for Pareto(1,*ξ*) for varying* ξ. Left: *Exponentiated outliers.* Middle: *Scaled outliers.* Right: *Mixed outliers.*

4.5. Outliers in non Pareto distributions

In this section, $n = 1000$ sample points are generated from non-Pareto distributions as in Relation (4.2). These include the |T|(ξ) and the Burr(η,λ,ξ) distribution with $\xi = 2$, $\eta = 1$ and $\lambda = 0.5$. Outliers are injected by mechanisms (4.4), (4.5) and (4.6) with $L = 3$, $C = 200$, $M = 100$, $k_0 = 10$ and $\tau = 5000$. The adaptive trimmed Hill estimator (ADAP) is constructed with k in the neighborhood of its optimal value as in Relation (4.3)[2].

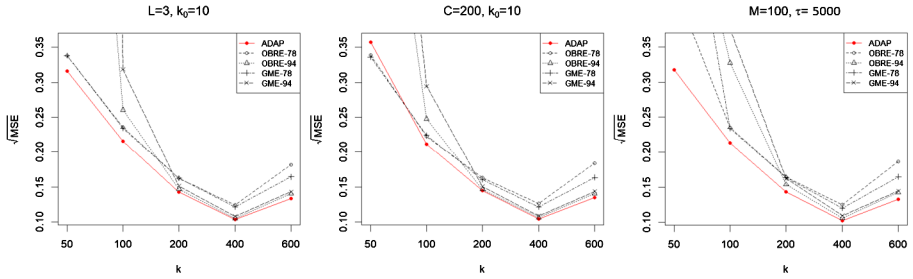

FIG 9. $\sqrt{MSE}$ *of ADAP for* $|T|(2)$ *as a function of* k. Left*: Exponentiated Outliers.* Middle*: Scaled Outliers.* Right*: Mixed Outliers.*

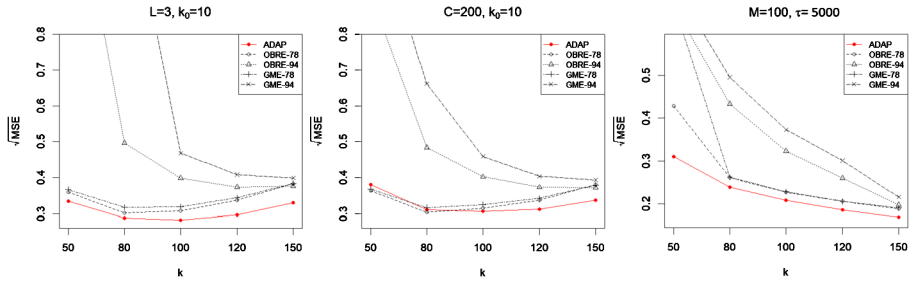

FIG 10. $\sqrt{MSE}$ *of adaptive trimmed Hill for Burr(1,0.5,2) as a function of* k. Left*: Exponentiated Outliers.* Middle*: Scaled Outliers.* Right*: Mixed Outliers.*

In Tables 4, 5, 6 and 7, the exponentiated and scaled outliers have been referred to as exp and scl outliers respectively.

Figures 9 and 10 display the performance of the adaptive trimmed Hill estimator (ADAP) for |T|(2) and Burr(1,0.5,2), distributions, respectively together with that of the optimal B-robust estimator (OBRE) and the generalized median estimator (GME). Overall, the ADAP is uniformly better than the OBRE and the GME. Exceptions include small values of k for the scaled outliers. For $k < 100$, the OBRE-94 and the GME-94 break down completely irrespective

[2]The optimal k for the trimmed Hill estimator is of the same order as that of the classic Hill estimator (see Theorem 3.8). For a sample of size $n = 1000$, the optimal k is 464 and 97 for $|T|$ and Burr distributions, respectively.

of the nature of the outliers and distribution under study. The OBRE-78 and the GME-78, though more robust than the OBRE-94 and the GME-94, cannot surpass the efficiency of the ADAP. Also for mixed outliers, even the OBRE-78 and the GME-78 break down for $k = 50$. This is because the OBRE and the GME are immune to outliers only if their target ARE value is less than the ratio $1 - k_0/k$. This is another manifestation of the fact that the OBRE and the GME, unlike ADAP are not adaptive to the unknown levels of contamination in the extremes (see also Figures 4, 5 and 6 in Section 4.3).

TABLE 6
$\mathbb{E}(\widehat{k}_0) \pm$ Standard Error$(\widehat{k}_0)$ *for* $|T|(2)$ *for* $L = 3$, $C = 200$ *and* $k_0 = 10$

k	$k = 50$	$k = 100$	$k = 200$	$k = 400$	$k = 600$
Exp outliers	10.04 ± 0.91	10.01 ± 0.66	10.02 ± 0.72	10.02 ± 0.82	10.02 ± 0.78
Scl outliers	9.74 ± 1.71	9.86 ± 1.44	9.91 ± 1.00	9.91 ± 0.95	9.87 ± 0.98

TABLE 7
$\mathbb{E}(\widehat{k}_0) \pm$ Standard Error$(\widehat{k}_0)$ *for Burr(1,0.5,2) distribution for for* $L = 3$, $C = 200$ *and* $k_0 = 10$

k	$k = 50$	$k = 80$	$k = 80$	$k = 100$	$k = 120$
Exp outliers	10.01 ± 0.71	10.00 ± 0.67	10.01 ± 0.88	9.98 ± 0.43	9.99 ± 0.49
Scl outliers	9.69 ± 1.72	9.76 ± 1.58	9.76 ± 1.36	9.78 ± 1.46	9.77 ± 1.29

Due to their slow rate convergence to Pareto tails, both Burr and |T| are difficult cases to analyze. For the Burr distribution with $\rho = 1$, the rate of convergence is further slower than that of the |T| with $\rho = 2\xi = 4$. However, the ADAP performs well even in this challenging regime. This can be attributed to the accuracy of the weighted sequential testing algorithm, Algorithm 1 which correctly identifies true number of outliers k_0 irrespective of the distribution under study for a wide range of k-values (see Tables 6 and 7).

5. Application

In this section, we apply our weighted sequential testing algorithm, Algorithm 1 and adaptive trimmed Hill estimator to real data. Two data sets have been explored in this context (see also the R shiny app at [29]). The first one provides the calcium content in the Condroz region of Belgium [35]. The data is indeed heavy tailed and has already been explored in the works of [5] and [37]. The second data set involves insurance claim settlements [9]. Both these data sets on analysis revealed the presence of outliers in the extremes and are therefore suitable for the application of our methodology.

5.1. Condroz data set

Figure 1 produces exploratory plots for the *Condroz data set* of [35] which measures the calcium content of soil samples together with their pH levels in the Condroz region of Belgium. As in [37], the conditional distribution of the

calcium content for pH levels lying between 7-7.5 has been considered. The left and middle panels use the value of $k = 85$ based on the k_{opt} value from [37]. The left panel displays a pareto quantile plot [5] of the data where an apparent linear trend indicates Pareto distributed observations. Nearly six data points show up as outliers in the pareto quantile plot. This has already been observed in [35] but no principled methodology for the identification of such outliers has been proposed. Our trimmed Hill estimator (recall Relation (2.2)) diagnostic plot in the middle panel also shows a change point in the values of the trimmed Hill statistics at $k_0 = 6$. On applying the Algorithm 1 with Type I error $q = 0.05$ and $a = 1.2$, we formally identify exactly $k_0 = 6$ outliers for this data set[3]. This is in consensus with the findings of [35] and [37].

The right panel in Figure 1 displays the values trimmed Hill estimator as a function of k for $k_0 = \widehat{k}_0 = 6$. Also displayed as a function of k are the values of the estimators, classic Hill and biased Hill with $k_0 = 6$ (recall Relations (1.2) and (1.4)). The robust estimator of ξ as reported in the analysis of [37] is same as that of the biased Hill. When compared with the trimmed Hill, the classic Hill plot produces much larger estimates and the biased Hill plot produces much smaller estimates of the tail index ξ. This can be explained by the apparent upward trend in the outliers as shown in left and middle panels of Figure 1. Thus, ignoring the presence of outliers by either using the classic Hill estimator or by naively truncating them and using the biased Hill statistic can lead to large discrepancies in the tail index values. The trimmed Hill estimator with $\widehat{k}_0 = 6$, also the adaptive trimmed Hill estimator, produces more credible estimates of the tail index ξ.

5.2. French claims data set

Next, we consider a data set of claim settlements issued by a private insurer in France for the time period 1996-2006 from [9]. We investigate the payments of claim settlements for the year 2006. Figure 11 produces exploratory plots of this data where the left and middle panels use the value of $k = 130$. The left panel displays a pareto quantile plot [5] of the data where an apparent linear trend indicates Pareto distributed observations as well as a large number of outliers. Nearly thirty three data points show up as outliers in the pareto quantile plot. This is further confirmed by the diagnostic plot in the middle panel where a change point in the values of trimmed Hill statistics is evident at $k_0 \approx 33$. On applying the Algorithm 1 with $q = 0.05$ and $a = 1.2$, we identify $k_0 \approx 33$ outliers for this data set.

In contrast to the case of Condroz data set (Figure 1 right panel), now the both classic and biased Hill plots lie under the trimmed Hill plot (see the right panel of Figure 11 constructed with $k_0 = 33$ and varying k). This can be explained by the apparent downward trend in the outliers as shown in left and middle panels of Figure 11.

[3]The ties in the data are broken using a suitable dithering technique like adding a small perturbation $\epsilon \sim U(0, 0.1)$ to the data or considering unique values in the data

Observe that the trimmed Hill plot in Figure 11 (right panel) has a rather high peak for k close to k_0, but then it quickly stabilizes around the value of 2, when k grows. It is well-known that except in the ideal Pareto setting, the classic Hill plot can be quite volatile for small values of k (see Figure 4.2 in [34]). The same holds for the trimmed Hill plots, but ultimately, in Figure 11 for a wide range of k's the trimmed Hill plot is relatively stable and it provides more reliable estimates of ξ than the classic and biased Hill plots therein. This simple analysis shows that ignoring or not adequately treating extreme outliers can lead to significant underestimation of the tail index ξ. This in turn can result in severe underestimation of the tail of loss distribution with detrimental effects to the insurance industry.

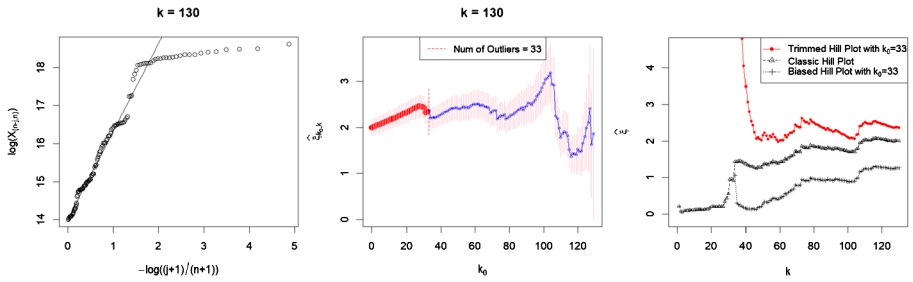

FIG 11. *Exploratory plots of the French claim settlements.* Left: *Pareto quantile plot,* Middle: *Diagnostic Plot and* Right: *Hill plots viz classic Hill plot, trimmed Hill plot and biased Hill plot.*

Appendix A

A.1. Empirical estimation of the rates of convergence

If a statistic $\mathcal{T}_n$ satisfies

$$n^r \mathcal{T}_n = \mathcal{O}_{\mathbb{P}}(1), \quad \text{as } n \to \infty,$$

then the optimal (largest) r can be estimated empirically by numerical simulations. Adopting the simplifying assumptions that $n^r \mathcal{T}_n \to Z$, in probability, for some finite and possibly random Z and that $\{|n^r \mathcal{T}_n|,\ n \in \mathbb{N}\}$ are uniformly integrable, for example. Then, we will have $n^r m_n \to \mathbb{E}|Z| < \infty$, where $m_n := \mathbb{E}|\mathcal{T}_n|$. One can compute empirical Monte Carlo estimates $\widehat{m}_n$ of m_n using a large number of independent realizations of $\mathcal{T}_n$, for each n from a range of large sample sizes. Then, the negative of slope estimate in a log-linear regression of $\widehat{m}_n$ versus n can be taken as an estimate of the rate r. We employed this simple method below to gain some intuition behind the optimal rates for the statistics involved in our adaptive trimmed Hill estimator.

A.2. Rates of convergence of $\widehat{\xi}_{k_0,k}$, $T_{k_0,k}$ and $U_{k_0,k}$

In Theorem 3.5, it has been shown that whenever Relations (3.9) and (3.10) hold,

$$k^{\delta} \max_{0 \le k_0 < h(k)} \left| \widehat{\xi}_{k_0,k} - \widehat{\xi}^{*}_{k_0,k} - \frac{cAk^{-\delta}}{1+\rho} \right| \xrightarrow{\mathbb{P}} 0. \tag{A.1}$$

where $\widehat{\xi}_{k_0,k}$ and $\widehat{\xi^*}_{k_0,k}$ are defined in Relation (3.5). Whether the rate δ is optimal or not is what we explore next.

Simulation setting For the Hall class of distributions as in Relation (3.6), $g(x) \approx x^{-\rho}$ which implies Relation (3.10) holds if $k = k(n) \sim n^{\rho/(\rho+\delta)}$. We thereby work with this choice of k, for varying sample sizes n. We simulate a sample of size n from Pareto(1,1) and then construct a sample from a general distribution function F by using Relation (3.4). Then, $\widehat{\xi}_{k_0,k}$ is the trimmed Hill estimator based on the sample from F (see Relation (2.2)). Let ξ be the tail index associated with F, then $\widehat{\xi^*}_{k_0,k} = \xi\widehat{\xi}$, where $\widehat{\xi}$ is the trimmed Hill estimator based on Pareto(1,1) sample. Three different distribution functions F are considered viz |T|(0.25), Burr$(1, 0.5, 1)$ and |T|(2). The sample sizes used are $n = 10^5 \times \{1, 2, 5, 10, 20\}$, $n = 10^3 \times \{1, 2, 5, 10, 20\}$ and $n = 10^2 \times \{1, 2, 5, 10, 20\}$, for |T|(0.25), Burr$(1, 0.5, 1)$ and |T|(2), respectively.

Using Section A.1, we numerically evaluate the rate of convergence for $|\widehat{\xi}_{k_0,k} - \widehat{\xi}^{*}_{k_0,k}|$ for varying δ, n and F. The rate $\hat{\delta}_e$ so obtained is a power of n which is then converted to a power of k by noting that $n = k^{(\rho+\delta)/\rho}$. Table 8 gives the rates $\hat{\delta}_e$ as a power of k. The values thus obtained closely follow the true value of δ used in Relation (3.10), irrespective of the value of ρ and the distribution function F. Thus, the rate δ in Relation (A.1) cannot be improved further.

In Theorem 3.9, we prove that whenever Relations (3.9) and (3.10) hold,

$$k^{\delta} |T_{k_0,k} - T^{*}_{k_0,k}| \xrightarrow{\mathbb{P}} 0 \tag{A.2}$$

where $T_{k_0,k}$ and $T^{*}_{k_0,k}$ are defined using $\widehat{\xi}_{k_0,k}$ and $\widehat{\xi^*}_{k_0,k}$, respectively, based on Relation (2.6). Whether the rate δ is the best achievable rate or not is what we explore next.

TABLE 8

Rate of convergences. $\hat{\delta}_e$: convergence rate for $|\widehat{\xi}_{k_0,k} - \widehat{\xi}^{*}_{k_0,k}|$, $\hat{\delta}_t$: convergence rate for $|T_{k_0,k} - T^{*}_{k_0,k}|$ and $\hat{\delta}_u$: convergence rate for $|U_{k_0,k} - U^{*}_{k_0,k}|$

	|T|(0.25), $\rho = 0.5$			Burr$(1, 0.5, 1)$, $\rho = 1$			|T|(2), $\rho = 4$		
δ	$\hat{\delta}_e$	$\hat{\delta}_t$	$\hat{\delta}_t - \hat{\delta}_u$	$\hat{\delta}_e$	$\hat{\delta}_t$	$\hat{\delta}_t - \hat{\delta}_u$	$\hat{\delta}_e$	$\hat{\delta}_t$	$\hat{\delta}_t - \hat{\delta}_u$
0.5	0.57	1.34	0.88	0.51	1.36	0.86	0.51	1.17	0.64
0.75	0.78	1.54	0.86	0.75	1.58	0.84	0.75	1.42	0.62
1	1.07	1.73	0.83	0.99	1.78	0.8	0.99	1.64	0.59
1.25	1.34	1.87	0.79	1.23	2	0.78	1.26	1.9	0.56
1.5	1.66	2.08	0.75	1.47	2.21	0.75	1.53	2.16	0.53
2	2.35	2.62	0.66	2.01	2.68	0.69	2.08	2.68	0.49

We compute $T_{k_0,k}$ and $T^*_{k_0,k}$ based on $\widehat{\xi}_{k_0,k}$ and $\widehat{\xi^*}_{k_0,k}$. Using Section A.1, the rate, $\hat{\delta}_t$ of convergence of $|T_{k_0,k} - T^*_{k_0,k}|$ is obtained as a power of n which is then converted to a power of k using $n = k^{(\rho+\delta)/\rho}$. Table 8 gives the rates $\hat{\delta}_t$ as a power of k. The results show that the numerically obtained rate, $\hat{\delta}_t$ is indeed larger than δ and lies approximately in the interval $[\delta + 0.5, \delta + 0.8]$ depending on the value of ρ and the distribution function F. This suggests that there is a scope of improvement of the rate in Relation (A.2). The proof, however, may require additional assumptions on the slowly varying function $\mathcal{L}$ in Relation (3.2) and goes beyond the scope of the present paper.

In Theorem 3.11 part 1., it has been shown that if

$$k^{\delta^*} |T_{k_0,k} - T^*_{k_0,k}| \xrightarrow{\mathbb{P}} 0 \tag{A.3}$$

holds for some $\delta^* > 0$, then

$$k^{(\delta^*-1)} |U_{k_0,k} - U^*_{k_0,k}| \xrightarrow{\mathbb{P}} 0 \tag{A.4}$$

where $U_{k_0,k}$ and $U^*_{k_0,k}$ are defined using $T_{k_0,k}$ and $T^*_{k_0,k}$, respectively, based on Relation (2.7). Thus, we expect the rate of convergence of $|U_{k_0,k} - U^*_{k_0,k}|$ to differ from that of $|T_{k_0,k} - T^*_{k_0,k}|$ by a margin of 1. Whether this difference can be further improved or not is what we explore next.

The rate of convergence, $\hat{\delta}_t$ for $|T_{k_0,k} - T^*_{k_0,k}|$ has already been obtained. Using Section A.1, the rate, $\hat{\delta}_u$ of convergence of $|U_{k_0,k} - U^*_{k_0,k}|$ is obtained as a power of n which is then converted to a power of k by using $n = k^{(\rho+\delta)/\rho}$. Table 8 gives the values of the difference, $\hat{\delta}_t - \hat{\delta}_u$. The results demonstrate that $\hat{\delta}_t - \hat{\delta}_u$ lies approximately in the interval $[0.5, 0.9]$ depending on the value of ρ and the distribution function F. This indicates that the convergence rate of $|U_{k_0,k} - U^*_{k_0,k}|$ is nearly 1 unit smaller than the rate for $|T_{k_0,k} - T^*_{k_0,k}|$. Thus, for a given rate δ^* in Relation (A.3), the rate $\delta^* - 1$ in Relation (A.4) cannot be improved further.

A.3. Rate of convergence of Type I error

In Theorem 3.11 part 2., it has been proved that if Relation (A.2) holds for some $\delta^* > 0$, then under $\mathcal{H}_0 : k_0 = 0$,

$$P_{\mathcal{H}_0}[\widehat{k}_0 > 0] \longrightarrow q$$

as long as $\delta^* \geq 2$. Here $\widehat{k}_0$ is the output of weighted sequential testing, Algorithm 1 when applied to the statistics $U_{k_0,k}$ constructed from $T_{k_0,k}$ using Relation (2.7). The question of whether the condition $\delta^* \geq 2$ is necessary or not is explored next using the same simulation setting as in Section A.2.

For the computation $\widehat{k}_0$, the significance level q in Algorithm 1 is fixed at 0.05. The Type I error, $P_{\mathcal{H}_0}[\widehat{k}_0 > 0]$ is computed empirically by considering the proportion of Monte Carlo iterations where $\widehat{k}_0 > 0$. The rate of convergence, $\hat{\delta}_t$

for $|T_{k_0,k} - T^*_{k_0,k}|$ as obtained in Section A.2 is considered. For varying values of the rate ($\hat{\delta}_t$ is denoted as δ in Figure 12), Figure 12 gives a plot of the Type I error as a function of the sample size n. The plot clearly shows that as $\hat{\delta}_t$ approaches 2, the Type I error converges to the true significance level 0.05. For values of $\hat{\delta}_t$ smaller than 2, the Type I error is much larger than the significance level 0.05. This suggests that the condition $\delta^* \geq 2$ is necessary and cannot be further relaxed.

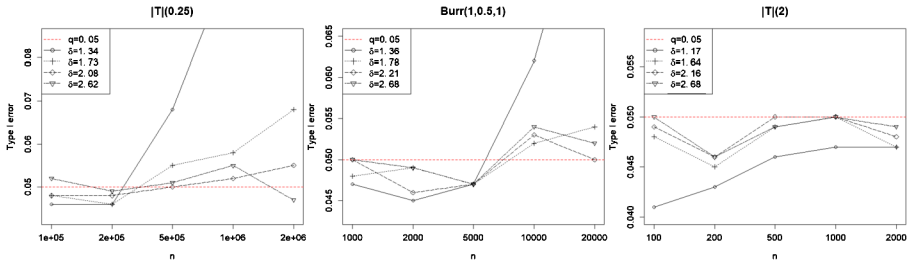

FIG 12. *Type I error at varying values of the rate of convergence for* $|T_{k_0,k} - T^*_{k_0,k}|$ *(denoted by* δ*).* Left: $|T|$ *with* $\rho = 0.5$, Middle: *Burr with* $\rho = 1$, Right: $|T|$ *with* $\rho = 4$.

Appendix B

B.1. The optimal B-robust estimator

The optimal B-robust estimator (OBRE) was first defined in [22] in terms of the influence function (IF), to allow for the assessment of robustness of an estimator in a parametric model. In [38], the OBRE estimator was adapted to the Pareto model to provide a robust estimator of the tail index ξ.

For a parametric model F_θ with density f_θ, $\theta \in \Theta \subseteq \mathbb{R}^p$, suppose $\mathcal{T}_n = \mathcal{T}_n(x_1, x_2, \cdots, x_n)$ is an estimator of θ for a sample of n observations, $x_1, ..., x_n$, from F_θ. Let F_n denote the empirical distribution function based on x_i's. The influence function (IF) viewed as a functional of F_n, $\mathcal{T}(F_n) = \mathcal{T}_n(x_1, x_2, \cdots, x_n)$ is defined as

$$IF(x; \mathcal{T}; F_\theta) = \lim_{\varepsilon \to 0} \frac{\mathcal{T}((1-\varepsilon)F_\theta + \varepsilon\delta_x) - \mathcal{T}(F_\theta)}{\varepsilon} \tag{B.1}$$

Thus, the IF describes the effect of a small contamination $\varepsilon\delta_x$ at the point x on the estimate standardized by the mass of the contamination.

An M-estimator $\mathcal{T}_n$ of θ satisfies

$$\sum_{i=1}^{n} \Psi(x_i, \mathcal{T}_n) = 0$$

for some function $\Psi : X \times \mathbb{R}^p \to \mathbb{R}^p$. Using (B.1), it can be shown that the IF of an M-estimator defined by Ψ at F_θ is given by

$$IF(x; \Psi, F_\theta) = \Big[-\int \frac{\delta}{\delta\theta} \Psi(x, \theta) dF_\theta(x) \Big]^{-1} \Psi(x, \theta)$$

The OBRE is an M-estimator which minimizes the trace of the asymptotic covariance matrix under the constraint that it has a bounded IF. For given bound $\widetilde{c}$ on the influence function (IF$\leq \widetilde{c}$), the OBRE is defined as the solution to the equation

$$\sum_{i=1}^{n} \Psi(x_i;\theta) = \sum_{i=1}^{n} (s(x_i;\theta) - a(\theta)) W_{\widetilde{c}}(x_i;\theta) = 0 \tag{B.2}$$

where

$$s(x,\theta) = \frac{\delta}{\delta\theta} \log f_\theta(x) \qquad W_{\widetilde{c}}(x;\theta) = \min\left(1, \frac{\widetilde{c}}{||A(\theta)(s(x;\theta) - a(\theta))||}\right). \tag{B.3}$$

The matrix $A(\theta)$ and the vector $a(\theta)$ in Relations (B.2) and (B.3) are defined implicitly by

$$E[\Psi(x;\theta)\Psi(x;\theta)^\top] = (A(\theta)^\top A(\theta))^{-1} \qquad E[\Psi(x;\theta)] = 0$$

The constant $\widetilde{c}$ may be interpreted as the regulator between robustness and efficiency, wherein for a lower $\widetilde{c}$ one gains robustness but loses efficiency, and vice versa for a higher $\widetilde{c}$.

For $\widetilde{c} = \infty$, one obtains the MLE. For a given $\widetilde{c}$, the asymptotic relative efficiency of the OBRE (expressed as a proportion) is defined as the ratio of the traces of the asymptotic covariance of the MLE to that of the OBRE. Its explicit form is given by

$$\frac{\mathrm{tr}\{(\int s(x;\theta)s(x;\theta)^\top dF_\theta(x))^{-1}\}}{\mathrm{tr}\{\int IF(x,\Psi,F_\theta)IF(x,\Psi,F_\theta)^\top dF_\theta(x)dx\}}.$$

where Ψ has the form as in (B.2).

For the Pareto(σ,θ) model, the distribution function F_θ is given by Relation (4.2). It so turns out that the OBRE based on Relation (B.2) for Pareto(σ,θ) has ARE values equal to 78% and 94% for $\widetilde{c} = 1.63$ and 2.73, respectively.

B.2. The generalized median estimator

The generalized median estimator (GME) is another robust estimator of the tail index developed by [12] for Pareto models. Let $X \sim \mathrm{Pareto}(\sigma,\theta)$ (see Relation (4.2)), then $Z = \log X$ is exponentially distributed with location and scale parameters $\log\sigma$ and θ, respectively. The cumulative distribution function (cdf) for exponential distribution with location and scale μ and ζ, respectively (denoted by $\mathrm{Exp}(\mu,\zeta)$) has the form

$$G(x) = 1 - e^{-(x-\mu)/\zeta}. \tag{B.4}$$

For a sample $x_1, x_2, \cdots, x_n$ from Pareto(σ,θ), the generalized median estimator seeks to obtain an estimator of θ by using the sample $z_i = \log x_i$, $i = 1, 2, \cdots, n$ from Exp($\log\sigma,\theta$).

For a kernel $h(z_1, z_2, \cdots, z_\kappa)$ invariant under permutation of its κ arguments, let H_F denote the induced cdf of $h(Z_1, Z_2, \cdots, Z_\kappa)$ where F is the cdf of $Z_i \sim \text{Exp}(\log\sigma, \theta)$. The kernel h is chosen such that θ is the median of H_F as follows

$$\theta = H_F^{-1}(0.5) \tag{B.5}$$

Motivated by Relation (B.5), the generalized median estimator (GME) of θ as defined in [12] is:

$$\widehat{\theta}_{\text{GM}} = \widehat{H}_n^{-1}(0.5) \tag{B.6}$$

where $\widehat{H}_n$, an estimator of H_F and has the form

$$\widehat{H}_n(y) = \frac{1}{\binom{n}{\kappa}} \sum \mathbf{1}[h(z_{i1}, z_{i2}, \cdots, z_{i\kappa}) \le y], \quad y \in \mathbb{R}. \tag{B.7}$$

Here, the sum is over all possible κ-sets of distinct indices $\{i_1, i_2, \cdots, i_\kappa\}$ from $\{1, 2, \cdots, n\}$.

The asymptotic relative efficiency (ARE) of GME (expressed as a proportion) is the ratio of the variances of GME and MLE of θ and is given by

$$\frac{n h_F^2(\theta)}{v\kappa^2}$$

where h_F is the density of H_F and $v = \text{Var}(w_h(Z))$ for $w_h(z) = P(h(z, Z_1, \cdots, Z_{\kappa-1}) \le \theta)$.

Two forms of the kernel h have been proposed in [12] of which we use the one with higher ARE. The form of this kernel h is given by

$$h(z_1, z_2, \cdots, z_\kappa) = \frac{2\kappa}{M_{2\kappa,0.5}} \Big(\frac{1}{\kappa} \sum_{j=1}^{\kappa} z_j - z_{(1,n)} \Big) \tag{B.8}$$

where $z_{1,n} = \min(z_1, z_2, \cdots, z_n)$ and $M_{2\kappa,0.5}$ is the median (0.5^{th} quantile) for $\chi^2_{2\kappa}$ distribution. For smaller sample sizes, $M_{2\kappa,0.5}$ is replaced by $\widetilde{M}_{2\kappa,0.5}$ where $\widetilde{M}_{2\kappa,0.5}$ is the median (0.5^{th} quantile) of the mixture distribution

$$\Big(1 - \frac{\kappa}{n}\Big)\chi^2_{2\kappa} + \Big(\frac{\kappa}{n}\Big)\chi^2_{2(\kappa-1)}.$$

The constant κ serves as a regulator between the robustness and ARE. Indeed with h as in Relation (B.8), values of $\kappa = 2$ and $\kappa = 5$ produce ARE levels equal to 78% and 94%, respectively.

Appendix C

C.1. Auxiliary lemmas

Lemma C.1. *Let E_j, $j = 1, 2, \cdots, n+1$ be i.i.d. standard exponential random variables. Then, the* Gamma$(i, 1)$ *random variables defined as*

$$\Gamma_i = \sum_{j=1}^{i} E_j \quad i = 1, \cdots, n+1, \tag{C.1}$$

satisfy

$$\Big(\frac{\Gamma_1}{\Gamma_{n+1}},\cdots,\frac{\Gamma_n}{\Gamma_{n+1}}\Big) \text{ and } \Gamma_{n+1} \text{ are independent.} \tag{C.2}$$

and

$$\Big(\frac{\Gamma_1}{\Gamma_{n+1}},\cdots,\frac{\Gamma_n}{\Gamma_{n+1}}\Big) \stackrel{d}{=} (U_{(1,n)},\cdots,U_{(n,n)}) \tag{C.3}$$

where $U_{(1,n)} < \cdots < U_{(n,n)}$ are the order statistics of n i.i.d. U(0,1) random variables.

For details on the proof see Example 4.6 on page 44 in [3].

The next result is used throughout the course of the paper to switch between order statistics of exponentials and i.i.d. exponential random variables.

Lemma C.2 (Rényi, 1953 [15])**.** *Let $E_1, E_2, \cdots, E_n$ be a sample of n i.i.d. exponential random variables with mean ξ (denoted by $\mathrm{Exp}(\xi)$) and $E_{(1,n)} \le E_{(2,n)} \le E_{(n,n)}$ be the order statistics. By Rényi's (1953) representation on page 37 of [15], we have for fixed $k \le n$,*

$$(E_{(1,n)},\cdots,E_{(i,n)},\cdots,E_{(k,n)}) \stackrel{d}{=} \Big(\frac{E_1^*}{n},\cdots,\sum_{j=1}^{i}\frac{E_j^*}{n-j+1},\cdots,\sum_{j=1}^{k}\frac{E_j^*}{n-j+1}\Big) \tag{C.4}$$

where $E_1^*,\cdots,E_k^*$ are also i.i.d. $\mathrm{Exp}(\xi)$.

Lemma C.3. *For $\Gamma_m = E_1 + E_2 + \cdots + E_m$ where the $E_i's$ i.i.d. $\mathrm{Exp}(\xi)$, then for any $\rho \in (-\infty,\infty)$*

$$\sup_{m\ge M}\Big|\Big(\frac{\Gamma_m}{m}\Big)^{\rho}-1\Big| \stackrel{a.s.}{\longrightarrow} 0, \quad M\to\infty. \tag{C.5}$$

$$\sup_{m,n\ge M}\Big|\Big(\frac{\Gamma_m/m}{\Gamma_n/n}\Big)^{\rho}-1\Big| \stackrel{a.s.}{\longrightarrow} 0, \quad M\to\infty. \tag{C.6}$$

Lemma C.4. *For the setup of Lemma C.3, for all $\rho \in [0,\infty)$, we have*

$$\sup_{m\ge M}\Big|\frac{1}{m}\sum_{i=1}^{m}\Big(\frac{\Gamma_{i+1}}{\Gamma_{m+1}}\Big)^{\rho}-\frac{1}{1+\rho}\Big| \stackrel{a.s.}{\longrightarrow} 0, \quad M\to\infty$$

Proof. It is equivalent to show that, as $m\to\infty$,

$$\Big|\frac{1}{m}\sum_{i=1}^{m}\Big(\frac{\Gamma_{i+1}}{\Gamma_{m+1}}\Big)^{\rho}-\frac{1}{1+\rho}\Big| \stackrel{a.s.}{\longrightarrow} 0. \tag{C.7}$$

For a fixed $\omega\in\Omega$, let us define the following sequence of functions

$$f_m(x) = \sum_{i=1}^{m}(\Gamma_{i+1}/\Gamma_{m+1})^{\rho}(\omega)\mathbf{1}_{(\frac{i-1}{m},\frac{i}{m}]}(x), \qquad x>0$$

Suppose $x \in ((i-1)/m, i/m]$, then

$$f_m(x) = (\Gamma_{[mx]+1}/\Gamma_{m+1})^\rho(\omega) = \Big(\frac{[mx]+1}{m}\Big)^\rho \Big(\frac{\Gamma_{[mx]+1}/([mx]+1)}{\Gamma_m/m}\Big)^\rho(\omega) \to x^\rho \tag{C.8}$$

where the convergence follows from Relation (C.6). Moreover since $\Gamma_{[mx]+1} < \Gamma_m$ and $\rho \geq 0$, therefore $|f_m(x)| \leq 1$, for all $x > 0$. Thus by dominated convergence theorem,

$$\int_0^1 f_m(x)dx = \frac{1}{m}\sum_{i=1}^m (\Gamma_{i+1}/\Gamma_{m+1})^\rho(\omega) \to \int_0^1 x^\rho dx = \frac{1}{1+\rho} \tag{C.9}$$

Since Relation (C.8) holds for all $\omega \in \Omega$ with $\mathbb{P}(\Omega] = 1$, so does Relation (C.9). This completes the proof. □

C.2. Proofs for Section 2

Proof of Proposition 2.1. Note that, if $X_i \sim \text{Pareto}(\sigma, \xi)$, then it can be alternatively written as

$$X_i = \sigma U_i^{-\xi}, \quad i = 1, \cdots, n,$$

where U_i's are i.i.d. $U(0,1)$. Therefore by Relation (C.3), we have

$$(X_{(n,n)}, \cdots, X_{(1,n)}) = \sigma(U_{(1,n)}^{-\xi}, \cdots, U_{(n,n)}^{-\xi}) \stackrel{d}{=} \sigma\left(\Big(\frac{\Gamma_1}{\Gamma_{n+1}}\Big)^{-\xi}, \cdots, \Big(\frac{\Gamma_n}{\Gamma_{n+1}}\Big)^{-\xi}\right) \tag{C.10}$$

where $X_{(n,n)} > \cdots > X_{(1,n)}$ are the order statistics for the X_i's. Hence, for all $1 \leq k \leq n-1$, we have

$$\begin{aligned} &\left(\log\Big(\frac{X_{(n,n)}}{X_{(n-k,n)}}\Big), \cdots, \log\Big(\frac{X_{(n-k+1,n)}}{X_{(n-k,n)}}\Big)\right) \\ \stackrel{d}{=}\ & -\xi\left(\log\Big(\frac{\Gamma_1}{\Gamma_{k+1}}\Big), \cdots, \log\Big(\frac{\Gamma_k}{\Gamma_{k+1}}\Big)\right) \\ \stackrel{d}{=}\ & -\xi(\log U_{(1,k)}, \cdots, \log U_{(k,k)}), \end{aligned} \tag{C.11}$$

where the $U_{(i,k)}$'s are the order statistics for a sample of k i.i.d. $U(0,1)$ and the last equality in Relation (C.11) follows from Relation (C.3). Since negative log transforms of $U(0,1)$ are standard exponentials, one can define $E_{(i,k)}$, $i = 1, \cdots, k$ as

$$\left(\log\Big(\frac{X_{(n,n)}}{X_{(n-k,n)}}\Big), \cdots, \log\Big(\frac{X_{(n-k+1,n)}}{X_{(n-k,n)}}\Big)\right) =: (E_{(k,k)}, \cdots, E_{(1,k)}) \tag{C.12}$$

such that the $E_{(i,k)}$'s are the order statistics of k i.i.d. exponentials with mean ξ.

Using Relation (C.12), $\widehat{\xi}^{\text{trim}}_{k_0,k}$ in Relation (1.3) is simplified as:

$$\widehat{\xi}^{\text{trim}}_{k_0,k} = \sum_{i=k_0+1}^{k} c_{k_0,k}(i) E_{(k-i+1,k)} = \sum_{i=1}^{k-k_0} \delta_i E_{(i,k)} \tag{C.13}$$

where $\delta_i = c_{k_0,k}(k-i+1)$. The optimal choice of weights δ_i's which produces the best linear unbiased estimator (BLUE) is obtained using Lemma C.5 below as follows:

$$\delta_i^{\text{opt}} = \begin{cases} \frac{1}{k-k_0} & i = 1, \cdots, k-k_0-1 \\ \frac{k_0+1}{k-k_0} & i = k-k_0 \end{cases} \tag{C.14}$$

Rewriting $E_{(i,k)}$'s in terms of $X_{(n-i+1,n)}$'s as in Relation (C.12) completes the proof. □

Lemma C.5. *If E_i, $i = 1, \cdots, n$ are i.i.d. observations from* Exp(ξ)*, the best linear unbiased estimator (BLUE) of ξ based on the order statistics, $E_{(1,n)} < \cdots < E_{(r,n)}$ is given by*

$$\widehat{\xi} = \frac{1}{r} \sum_{i=1}^{r-1} E_{(i,n)} + \frac{n-r+1}{r} E_{(r,n)}$$

Proof. Let $\widehat{\xi} = \sum_{i=1}^{r} \gamma_i E_{(i,n)}$ denote the BLUE of ξ. By Relation (C.4), the BLUE can then be expressed as

$$\widehat{\xi} = \sum_{i=1}^{r} \gamma_i \sum_{j=1}^{i} \frac{E_j^*}{(n-j+1)} = \sum_{j=1}^{r} E_j^* \sum_{i=j}^{r} \frac{\gamma_i}{(n-j+1)} =: \sum_{j=1}^{r} E_j^* \delta_j \tag{C.15}$$

where the E_j^* are i.i.d. from Exp(ξ) and $\delta_j = (n-j+1)^{-1} \sum_{i=j}^{r} \gamma_i$

For i.i.d. observations from Exp(ξ), the sample mean is the uniformly minimum variance unbiased estimator (UMVUE) for ξ (see Lehmann Scheffe Theorem, Theorem 1.11, page 88 in [31]).

Thus, $\delta_j = 1/r$ yields the required best linear unbiased estimator and therefore, the weights γ_i's have the form:

$$\gamma_i = \begin{cases} \frac{n-r+1}{r} & i = r \\ \frac{1}{r} & i < r \end{cases}$$

This completes the proof. □

Proof of Theorem 2.5. Assume that σ is known and consider the class of statistics:

$$\mathcal{U}_{k_0}^{\sigma} = \left\{ \mathcal{T} = \mathcal{T}(X_{(n-k_0,n)}, \cdots, X_{(1,n)}) : \mathbb{E}(\mathcal{T}) = \xi, X_1, \cdots, X_n \overset{i.i.d.}{\sim} \text{Pareto}(\sigma, \xi) \right\}.$$

Since σ is no longer a parameter, every statistic in $\mathcal{U}_{k_0}^{\sigma}$ can be equivalently written as a function of $\log(X_{(n-i+1,n)}/\sigma)$, $i = k_0+1, \cdots, n$ as follows:

$$\begin{aligned}\mathcal{U}_{k_0}^{\sigma} &= \Bigg\{\mathcal{S} = \mathcal{S}\left(\log\left(\frac{X_{(n-k_0,n)}}{\sigma}\right), \cdots, \log\left(\frac{X_{(1,n)}}{\sigma}\right)\right) \\ &: \mathbb{E}(\mathcal{S}) = \xi,\ X_1, \cdots, X_n \overset{i.i.d.}{\sim} \text{Pareto}(\sigma, \xi)\Bigg\}\end{aligned}$$

Since X_i's follow Pareto(σ, ξ), $\log(X_i/\sigma) \sim \text{Exp}(\xi)$ and therefore

$$\left(\log\left(\frac{X_{(n-k_0,n)}}{\sigma}\right), \cdots, \log\left(\frac{X_{(1,n)}}{\sigma}\right)\right) \overset{d}{=} \left(E_{(n-k_0,n)}, \cdots, E_{(1,n)}\right),$$

where $E_{(1,n)} \leq \cdots \leq E_{(n,n)}$ are the order statistics of n i.i.d. observations from Exp(ξ). Therefore

$$\mathcal{U}_{k_0}^{\sigma} \overset{d}{=} \left\{\mathcal{S} = \mathcal{S}(E_{(n-k_0,n)}, \cdots, E_{(1,n)}) : \mathbb{E}(\mathcal{S}) = \xi,\ E_1, \cdots, E_n \overset{i.i.d.}{\sim} \text{Exp}(\xi)\right\}, \tag{C.16}$$

where the E_i's do not depend on σ. Next, using Relation (C.4), we have

$$\begin{aligned}\mathcal{S}(E_{(n-k_0,n)}, \cdots, E_{(1,n)}) &= \mathcal{S}\Big(\sum_{j=1}^{n-k_0} \frac{E_j^*}{n-j+1}, \cdots, \sum_{j=1}^{n-k} \frac{E_j^*}{n-j+1}\Big) \\ &= \mathcal{R}(E_1^*, \cdots, E_{n-k_0}^*)\end{aligned}$$

Using the above result with Relation (C.16), we get

$$\mathcal{U}_{k_0}^{\sigma} \overset{d}{=} \mathcal{V}_{k_0} := \left\{\mathcal{R} = \mathcal{R}(E_1^*, \cdots, E_{n-k_0}^*) : \mathbb{E}(\mathcal{R}) = \xi, E_1^*, \cdots, E_{n-k_0}^* \overset{i.i.d.}{\sim} \text{Exp}(\xi)\right\} \tag{C.17}$$

where the first equality is in the sense of finite dimensional distributions.

By Relation (C.17), we have $\inf_{\mathcal{T} \in \mathcal{U}_{k_0}^{\sigma}} \text{Var}(\mathcal{T}) = \inf_{\mathcal{R} \in \mathcal{V}_{k_0}} \text{Var}(\mathcal{R}) := L^*$. Since the sample mean, $\overline{E}_{n-k_0}^* = \sum_{i=1}^{n-k_0} E_i^*/(n-k_0)$ is uniformly the minimum variance estimator (UMVUE) of ξ among the class described by $\mathcal{V}_{k_0}$, L^* can be easily obtained as

$$L^* = \text{Var}(\overline{E}_{n-k_0}^*) = \frac{\xi^2}{n-k_0} \tag{C.18}$$

The fact that $\overline{E}_{n-k_0}^*$ is the UMVUE follows because it is an unbiased and complete sufficient statistic for ξ (see Lehmann Scheffe Theorem, Theorem 1.11, page 88 in [31]).

To complete the proof, observe that every statistic $\mathcal{T}$ in $\mathcal{U}_{k_0}$ is an unbiased estimator of ξ for any arbitrary choice of σ. This implies that for any σ, $\mathcal{T} \in \mathcal{U}_{k_0}^{\sigma}$ and therefore $L^* \leq \text{Var}(\mathcal{T})$. Since this holds for all values of $\mathcal{T} \in \mathcal{U}_{k_0}$, the proof of the lower bound in Relation (2.3) follows.

For the upper bound in Relation (2.3), we observe that $\widehat{\xi}_{k_0,n-1} \in \mathcal{U}_{k_0}$, which in view of Proposition 2.6 implies

$$\inf_{\mathcal{T}\in\mathcal{U}_{k_0}} \mathrm{Var}(\mathcal{T}) \le \mathrm{Var}(\widehat{\xi}_{k_0,n-1}) = \frac{\xi^2}{n-k_0-1}.$$

This completes the proof. □

Proof of Proposition 2.6. From Relations (C.13) and (C.14), we have

$$\begin{aligned} &\left\{\widehat{\xi}_{k_0,k},\ k_0 = 0,\dots,k-1\right\} \qquad (C.19)\\ = &\left\{\frac{1}{k-k_0}\sum_{i=1}^{k-k_0-1} E_{(i,k)} + \frac{k_0+1}{k-k_0}E_{(k-k_0,k)},\ k_0 = 0,\dots,k-1\right\} \end{aligned}$$

Using Relation (C.4), for all $k_0 = 0, 1, \cdots, k-1$, we have

$$\widehat{\xi}_{k_0,k} = \frac{1}{k-k_0}\sum_{i=1}^{k-k_0-1}\sum_{j=1}^{i}\frac{E_j^*}{(k-j+1)} + \frac{k_0+1}{k-k_0}\sum_{j=1}^{k-k_0}\frac{E_j^*}{(k-j+1)} \qquad (C.20)$$

Interchanging the order of summation in the first term in the right hand side of Relation (C.20), we obtain

$$\begin{aligned} \widehat{\xi}_{k_0,k} &= \sum_{j=1}^{k-k_0-1}\frac{E_j^*}{k-j+1}\sum_{i=j}^{k-k_0-1}\frac{1}{k-k_0} + \frac{k_0+1}{k-k_0}\sum_{j=1}^{k-k_0}\frac{E_j^*}{(k-j+1)}\\ &= \sum_{j=1}^{k-k_0-1}\frac{E_j^*}{k-j+1}\left(\sum_{i=j}^{k-k_0-1}\frac{1}{k-k_0} + \frac{k_0+1}{k-k_0}\right) + \frac{E_{k-k_0}^*}{k-k_0}\\ &= \sum_{j=1}^{k-k_0-1}\frac{E_j^*}{k-j+1}\frac{(k-j+1)}{k-k_0} + \frac{E_{k-k_0}^*}{k-k_0}\\ &= \frac{1}{k-k_0}\sum_{j=1}^{k-k_0}E_j^*, \end{aligned}$$

Since E_j^*, $j = 1, \cdots, k-k_0$ follow $\mathrm{Exp}(\xi)$, E_j^* are indeed ξ times i.i.d. standard exponentials. This completes the proof of Relation (2.4).

The proof of Relation (2.5) is a direct application of central limit theorem to Relation (2.4). □

Proof of Proposition 2.7. In view of Relations (2.4) and (2.6), we have

$$\left(T_{0,k}(n),\cdots,T_{k-2,k}(n)\right) \stackrel{d}{=} \left(\frac{\Gamma_{k-1}}{\Gamma_k},\cdots,\frac{\Gamma_1}{\Gamma_2}\right), \qquad (C.21)$$

which implies

$$T_{k_0,k}(n) \stackrel{d}{=} \frac{\Gamma_{k-k_0-1}}{\Gamma_{k-k_0}} \sim \mathrm{Beta}(k-k_0-1,1), \qquad k_0 = 0,\cdots,k-2.$$

To show the independence of the $T_{k_0,k}(n)$'s, note that, by Relation (C.2) in Section C.1, Γ_m and $\{\Gamma_i/\Gamma_m, i = 1, \cdots, m\}$ are independent for all $1 \le m \le k-2$. This in turn implies that

$$\Big(\frac{\Gamma_1}{\Gamma_2}, \frac{\Gamma_2}{\Gamma_3}, \cdots, \frac{\Gamma_{m-1}}{\Gamma_m}\Big) \quad \text{and} \quad \Gamma_m \text{ are independent.}$$

Since Γ_i, $i = 1, \cdots, m$ and $(E_{m+1}, \cdots, E_k)$ are independent, for all $m = 1, \cdots, k-2$, we have

$$\Big(\frac{\Gamma_1}{\Gamma_2}, \cdots, \frac{\Gamma_{m-1}}{\Gamma_m}\Big) \quad \text{and} \quad (\Gamma_m, E_{m+1}, \cdots, E_k) \text{ are independent .} \tag{C.22}$$

Since $\Big(\Gamma_m/\Gamma_{m+1}, \cdots, \Gamma_{k-1}/\Gamma_k\Big)$ is a function of $(\Gamma_m, E_{m+1}, \cdots, E_k)$ for all $1 \le m \le k-2$, we have

$$\Big(\frac{\Gamma_1}{\Gamma_2}, \cdots, \frac{\Gamma_{m-1}}{\Gamma_m}\Big) \quad \text{and} \quad \Big(\frac{\Gamma_m}{\Gamma_{m+1}}, \cdots, \frac{\Gamma_{k-1}}{\Gamma_k}\Big) \text{ is independent for all } m \ge 1. \tag{C.23}$$

In view of Relations (C.21) and (C.23), the proof of independence of the $T_{k_0,k}(n)$'s follows. □

C.3. Proofs for Section 3

C.3.1. Minimax rate optimality

Our goal is to establish the uniform consistency in Relation (3.8). To this end, recall the representation in Relation (3.5). For the Hall class of distributions in Relation (3.6), it can be shown that $\sqrt{k}R_{k_0,k}$ is $O_{\mathbb{P}}(1)$ (see Lemma C.6 below). With $\sqrt{k}|R_{k_0,k}|$ bounded away from infinity, it is easier to bound the quantity $\sqrt{k}|\widehat{\xi}_{k_0,k} - \xi|$ since by Relation (3.5)

$$\sqrt{k}|\widehat{\xi}_{k_0,k} - \xi| \le \sqrt{k}|R_{k_0,k}| + \sqrt{k}|\widehat{\xi^*}_{k_0,k} - \xi|. \tag{C.24}$$

This shall form the basis of the proof for Theorem 3.3 as shown next.

Proof of Theorem 3.3. Let $P_n = \inf_{F \in \mathcal{D}_\xi(B,\rho)} \mathbb{P}_F\Big(\max_{0 \le k_0 < h(k)} |\widehat{\xi}_{k_0,k} - \xi| \le a(n)\Big)$. By Relation (C.24), we have

$$\begin{aligned} P_n \;=\; & \inf_{\mathcal{D}_\xi(B,\rho)} \mathbb{P}_F\Big(\underbrace{\max_{0 \le k_0 < h(k)} \sqrt{k}|R_{k_0,k}| \le (\sqrt{k}a(n))/2}_{A_{1n}} \\ & \cap \; \underbrace{\max_{0 \le k_0 < h(k)} \sqrt{k}|\widehat{\xi}^*_{k_0,k} - \xi| \le (\sqrt{k}a(n))/2}_{A_{2n}} \Big). \end{aligned}$$

Since $\sqrt{k}a(n) \to \infty$, therefore in view of Lemma C.6, $\inf_{F\in\mathcal{D}_\xi(B,\rho)} \mathbb{P}_F(A_{1n}) \to 1$. We also have that,

$$\inf_{F\in\mathcal{D}_\xi(B,\rho)} \mathbb{P}_F(A_{2n}) = \mathbb{P}\Big(\max_{0\le k_0<h(k)} \sqrt{k}|\widehat{\xi}^*_{k_0,k} - \xi| \le (\sqrt{k}a(n))/2\Big)$$

since $\widehat{\xi}^*_{k_0,k}$ does not depend on $F \in \mathcal{D}_\xi(B,\rho)$.

By using Donsker's principle, we will show that

$$\max_{0\le k_0<h(k)} |\widehat{\xi}^*_{k_0,k} - \xi| = o_{\mathbb{P}}(a(n)),$$

which will imply $\mathbb{P}_F(A_{2n}) \to 1$. Indeed, without loss of generality, suppose $\xi = 1$ and let $E_i,\ i = 1, 2, \dots$ be independent standard exponential random variables. For every $\epsilon \in (0, 1)$, we have that

$$W_k = \{W_k(t),\ t \in [\epsilon, 1]\} := \left\{ \frac{\sqrt{k}}{[kt]} \sum_{i=1}^{[kt]} (E_i - 1),\ t \in [0, 1] \right\} \xrightarrow{d} \{B(t)/t,\ t \in [\epsilon, 1]\}, \tag{C.25}$$

as $k \to \infty$, where $B = \{B(t),\ t \in [0, 1]\}$ is the standard Brownian motion, and where the last convergence is in the space of cadlag functions $\mathbb{D}[\epsilon, 1]$ equipped with the Skorokhod J_1-topology. (In fact, since the limit has continuous paths, the convergence is also valid in the uniform norm.)

Recall that by Relation (2.4), we have

$$\{\widehat{\xi}^*_{k_0,k}(n),\ 0 \le k_0 < k\} \stackrel{d}{=} \left\{ \sum_{i=1}^{k-k_0} E_i/(k - k_0),\ 0 \le k_0 < k \right\}.$$

Thus,

$$\sqrt{k} \max_{0\le k_0<h(k)} |\widehat{\xi}^*_{k_0,k}(n) - \xi| \stackrel{d}{=} \sup_{t\in[1-h(k)/k,1]} |W_k(t)| \le \sup_{t\in[\epsilon,1]} |W_k(t)|, \tag{C.26}$$

where the last inequality holds for all sufficiently large k, since $1 - h(k)/k \to 1$, as $k \to \infty$. Since the supremum is a continuous functional in J_1, the convergence in Relation (C.25) implies that the right–hand side of Relation (C.26) converges in distribution to $\sup_{t\in[\epsilon,1]} |B(t)/t| = O_{\mathbb{P}}(1)$, which is finite with probability one. This, since $a(n)\sqrt{k(n)} \to \infty$, completes the proof. □

Lemma C.6. *Assumption (3.6) implies there exist $M > 0$ such that*

$$\inf_{F\in\mathcal{D}_\xi(B,\rho)} \mathbb{P}_F\Big(\max_{0\le k_0<h(k)} \sqrt{k}|R_{k_0,k}| \le M\Big) \to 1 \text{ as } k \to \infty \tag{C.27}$$

where $R_{k_0,k}$ is defined as in Relation (3.5), $h(k) = o(k)$ and $k = O(n^{2\rho/(1+2\rho)})$.

Proof. By Relation (3.6), we have $1 - Bx^{-\rho} \le \mathcal{L}(x) \le 1 + Bx^{-\rho}$. Therefore,

$$\begin{aligned} & (k-k_0)R_{k_0,k} \qquad\qquad (C.28)\\ \le\ & (k_0+1)\log\frac{1+BY_{(n-k_0,n)}^{-\rho}}{1-BY_{(n-k,n)}^{-\rho}} + \sum_{i=k_0+2}^{k}\log\frac{1+BY_{(n-i+1,n)}^{-\rho}}{1-BY_{(n-k,n)}^{-\rho}} \\ \le\ & k\log\frac{1+BY_{(n-k,n)}^{-\rho}}{1-BY_{(n-k,n)}^{-\rho}}, \end{aligned}$$

since $Y_{(n-k,n)}^{-\rho} \ge Y_{(n-i+1,n)}^{-\rho}$ for $i = k_0+1, \cdots, k$. Similarly, we also have

$$(k-k_0)R_{k_0,k} \ge k\log\frac{1-BY_{(n-k,n)}^{-\rho}}{1+BY_{(n-k,n)}^{-\rho}} = -k\log\frac{1+BY_{(n-k,n)}^{-\rho}}{1-BY_{(n-k,n)}^{-\rho}}. \qquad (C.29)$$

Thus, Relations (C.28) and (C.29) together imply

$$\max_{0\le k_0<h(k)}\sqrt{k}|R_{k_0,k}| \le \frac{\sqrt{k}Y_{(n-k,n)}^{-\rho}}{1-h(k)/k}\max_{0\le k_0<h(k)}\frac{1}{Y_{(n-k,n)}^{-\rho}}\log\frac{1+BY_{(n-k,n)}^{-\rho}}{1-BY_{(n-k,n)}^{-\rho}} \qquad (C.30)$$

Since $h(k) = o(k)$, $1 - h(k)/k \to 1$. Additionally, expressing $Y_{(n-i+1,n)}$ in terms of Gamma random variables as in Relation (C.10), we get

$$\begin{aligned} & \sqrt{k}Y_{(n-k,n)}^{-\rho}\frac{1}{Y_{(n-k,n)}^{-\rho}}\log\frac{1+BY_{(n-k,n)}^{-\rho}}{1-BY_{(n-k,n)}^{-\rho}} \\ \overset{d}{=}\ & \underbrace{\sqrt{k}(\Gamma_{k+1}/\Gamma_{n+1})^{\rho}}_{\Delta_{1k}} \times \underbrace{\frac{1}{(\Gamma_{k+1}/\Gamma_{n+1})^{\rho}}\log\frac{1+B(\Gamma_{k+1}/\Gamma_{n+1})^{\rho}}{1-B(\Gamma_{k+1}/\Gamma_{n+1})^{\rho}}}_{\Delta_{2k}} \end{aligned}$$

Now, by Relation (C.5), we have $\Gamma_{k+1}/\Gamma_{n+1} \overset{a.s.}{\sim} (k/n)^{\rho}$. Therefore, for $k = O(n^{2\rho/(1+2\rho)})$, Δ_{1k} is $O_{\mathbb{P}}(1)$. Since $k/n \to 0$, therefore $(\Gamma_{k+1}/\Gamma_{n+1})^{\rho} \xrightarrow{a.s.} 0$ which further implies $\Delta_{2k} \xrightarrow{a.s.} 2B$ and is thereby $O_{\mathbb{P}}(1)$.

Thus, there exist M such that

$$\inf_{F\in\mathcal{D}_{\xi}(B,\rho)}\mathbb{P}_F\Big(\max_{0\le k_0<k}\frac{k-k_0}{kY_{(n-k,n)}^{-\rho}}|R_{k_0,k}| \le M\Big) \ge \mathbb{P}(\Delta_{1k}\Delta_{2k} \le M) \to 1$$

This completes the proof. □

C.3.2. Asymptotic normality

Proof of Theorem 3.5. To prove Relation (3.11), we observe that

$$k^{\delta}\Big|R_{k_0,k} - \frac{k^{-\delta}cA}{(1+\rho)}\Big| \le k^{\delta}|R_{k_0,k} - S_{k_0,k}| + k^{\delta}\Big|S_{k_0,k} - \frac{k^{-\delta}cA}{(1+\rho)}\Big| \qquad (C.31)$$

for $S_{k_0,k}$ defined as

$$\begin{aligned} S_{k_0,k} &:= \frac{cg(Y_{(n-k,n)})}{k-k_0}\Big((k_0+1)\int_1^{Y_{(n-k_0,n)}/Y_{(n-k,n)}} \nu^{-\rho-1}d\nu \\ &+ \sum_{i=k_0+2}^{k}\int_1^{Y_{(n-i+1,n)}/Y_{(n-k,n)}} \nu^{-\rho-1}d\nu\Big), \end{aligned} \tag{C.32}$$

where Y_i's are i.i.d observations from Pareto(1,1) as in (3.5).

We will show that the right hand side of (C.31) vanishes as $k \to \infty$. To this end, we first show that $k^\delta \max_{0\le k_0<h(k)} |R_{k_0,k} - S_{k_0,k}| \xrightarrow{\mathbb{P}} 0$ as follows:

$$\begin{aligned} & k^\delta \max_{0\le k_0<h(k)} |R_{k_0,k} - S_{k_0,k}| \\ = \; & k^\delta \max_{0\le k_0<h(k)} \frac{kg(Y_{n-k,n})}{k-k_0}\Big(\frac{k-k_0}{kg(Y_{(n-k,n)})}|R_{k_0,k} - S_{k_0,k}|\Big) \\ \le \; & \frac{k^\delta g(Y_{(n-k,n)})}{1-h(k)/k} \underbrace{\max_{0\le k_0<h(k)} \Big(\frac{k-k_0}{kg(Y_{(n-k,n)})}|R_{k_0,k} - S_{k_0,k}|\Big)}_{\Delta_{2k}} \end{aligned} \tag{C.33}$$

where $1-h(k)/k \to 1$ since $h(k) = o(k)$. Also, by Relation (C.40) and assumption (3.10),

$$k^\delta g(Y_{(n-k,n)}) \xrightarrow{\mathbb{P}} A \tag{C.34}$$

Thus, the convergence to 0 in probability of the last bound in Relation (C.33) follows from Lemma C.7, by which $\Delta_{2k} \xrightarrow{\mathbb{P}} 0$.

Next we show that the second term in the right hand side of (C.31) also vanishes. Indeed,

$$\begin{aligned} & k^\delta \max_{0\le k_0<h(k)} \Big|S_{k_0,k} - \frac{k^{-\delta}cA}{(1+\rho)}\Big| \\ = \; & k^\delta \max_{0\le k_0<h(k)} \frac{kg(Y_{n-k,n})}{k-k_0}\Big|\frac{k-k_0}{kg(Y_{(n-k,n)})}S_{k_0,k} - \frac{cA(k-k_0)}{(1+\rho)k^\delta g(Y_{(n-k,n)})}\Big| \\ \le \; & \frac{k^\delta g(Y_{(n-k,n)})}{1-h(k)/k} \underbrace{\max_{0\le k_0<h(k)} \Big|\frac{k-k_0}{kg(Y_{(n-k,n)})}S_{k_0,k} - \frac{cA(k-k_0)}{k(1+\rho)k^\delta g(Y_{(n-k,n)})}\Big|}_{\Delta_{3k}} \end{aligned}$$

where $k^\delta g(Y_{(n-k,n)}) \xrightarrow{\mathbb{P}} A$ as in Relation (C.34) and $1-h(k)/k \to 1$. Thus, the convergence to 0 in probability of the last upper bound follows because $\Delta_{3k} \xrightarrow{\mathbb{P}} 0$ as shown next.

$$\Delta_{3k} \le \underbrace{\max_{0\le k_0<h(k)} \Big|\frac{k-k_0}{kg(Y_{(n-k,n)})}S_{k_0,k} + c\Big(\frac{k_0}{k}\Big)^{1+\rho} - \frac{c}{1+\rho}\Big|}_{\Delta_{4k}}$$

$$+ \underbrace{\max_{0\le k_0<h(k)}\Big|\frac{c}{1+\rho}-c\Big(\frac{k_0}{k}\Big)^{1+\rho}-\frac{cA(k-k_0)}{k(1+\rho)k^{\delta}g(Y_{(n-k,n)})}\Big|}_{\Delta_{5k}}$$

where $\Delta_{4k} \xrightarrow{\mathbb{P}} 0$ by Lemma C.9. Next, we show that

$$\max_{0\le k_0<h(k)}\Big|\frac{c}{1+\rho}-\frac{cA(k-k_0)}{k(1+\rho)k^{\delta}g(Y_{(n-k,n)})}\Big| \xrightarrow{\mathbb{P}} 0$$

and since $\max_{0\le k_0<k}(k_0/k)^{1+\rho} \le (h(k)/k)^{1+\rho} \to 0$, the convergence to 0 in probability of Δ_{5k} shall follow. We have

$$\begin{aligned}
&\max_{0\le k_0\le h(k)}\Big|\frac{c}{1+\rho}-\frac{cA(k-k_0)}{k(1+\rho)k^{\delta}g(Y_{(n-k,n)})}\Big| \\
\le\ & \frac{|c|}{1+\rho}\max_{0\le k_0<h(k)}\left(\Big|1-\frac{A}{k^{\delta}g(Y_{(n-k,n)})}\Big|+\frac{Ak_0}{k^{\delta+1}g(Y_{(n-k,n)})}\right) \\
\le\ & \frac{|c|}{1+\rho}\left(\Big|1-\frac{A}{k^{\delta}g(Y_{(n-k,n)})}\Big|+\frac{Ah(k)}{k^{\delta+1}g(Y_{(n-k,n)})}\right)
\end{aligned}$$

where the last upper bound converges in probability to 0 because $h(k)/k \to 0$ and $A/k^{\delta}g(Y_{(n-k,n)}) \xrightarrow{\mathbb{P}} 1$ by Relation (C.34). This completes the proof. □

Lemma C.7. *Assumption (3.9) implies*

$$\max_{0\le k_0\le k}\Big(\frac{k-k_0}{kg(Y_{(n-k,n)})}|R_{k_0,k}-S_{k_0,k}|\Big) \xrightarrow{\mathbb{P}} 0 \tag{C.35}$$

where $R_{k_0,k}$ and $S_{k_0,k}$ are defined in Relations (3.5) and (C.32), respectively.

Proof. The proof of Relation (C.35) involves two cases: $\rho > 0$ and $\rho = 0$.

Case $\rho > 0$: Since $Y_{(n-i+1,n)}/Y_{(n-k,n)} > 1$, $i = 1, \cdots, k$, therefore, over the event $\{Y_{(n-k,n)} > t_\varepsilon\}$, by Relation (3.9), we have

$$\begin{aligned}
&(k-k_0)|R_{k_0,k}-S_{k_0,k}| \\
\le\ & (k_0+1)\left|\log\frac{\mathcal{L}(Y_{(n-k_0,n)})}{\mathcal{L}(Y_{(n-k,n)})}-cg(Y_{(n-k,n)})\int_1^{Y_{(n-k_0,n)}/Y_{(n-k,n)}}\nu^{-\rho-1}d\nu\right| \\
+\ & \sum_{i=k_0+2}^{k}\left|\log\frac{\mathcal{L}(Y_{(n-i+1,n)})}{\mathcal{L}(Y_{(n-k,n)})}-cg(Y_{(n-k,n)}\int_1^{Y_{(n-i+1,n)}/Y_{(n-k,n)}}\nu^{-\rho-1}d\nu\right| \\
\le\ & (k_0+1)g(Y_{(n-k,n)})\varepsilon+\sum_{i=k_0+2}^{k}g(Y_{(n-k,n)})\varepsilon = g(Y_{(n-k,n)})k\varepsilon.
\end{aligned}$$

Therefore, over the event $\{Y_{(n-k,n)} > t_\varepsilon\}$

$$\max_{0\le k_0\le k}\Big(\frac{k-k_0}{kg(Y_{(n-k,n)})}|R_{k_0,k}-S_{k_0,k}|\Big) \le \varepsilon. \tag{C.36}$$

From Relation (C.10), we have $Y_{(n-k,n)} \stackrel{d}{=} (\Gamma_{k+1}/\Gamma_{n+1})^{-1}$ where $(\Gamma_{k+1}/\Gamma_{n+1})^{-1} \stackrel{a.s.}{\sim} n/k$ by Lemma C.3. Since $n/k \to \infty$, therefore

$$\mathbb{P}(Y_{(n-k,n)} > t_\varepsilon) \to 1$$

which completes the proof.

Case $\rho = 0$*:* As in the previous case, over the event $\{Y_{(n-k,n)} > t_\varepsilon\}$, by Relation (3.9) we have

$$\begin{aligned}
& (k-k_0)|R_{k_0,k} - S_{k_0,k}| \\
= \;& (k_0+1)\left|\log \frac{\mathcal{L}(Y_{(n-k_0,n)})}{\mathcal{L}(Y_{(n-k,n)})} - cg(Y_{(n-k,n)}\int_1^{Y_{(n-k_0,n)}/Y_{(n-k,n)}} \frac{d\nu}{\nu}\right| \\
+ \;& \sum_{i=k_0+2}^{k}\left|\log \frac{\mathcal{L}(Y_{(n-i+1,n)})}{\mathcal{L}(Y_{(n-k,n)})} - cg(Y_{(n-k,n)}\int_1^{Y_{(n-i+1,n)}/Y_{(n-k,n)}} \frac{d\nu}{\nu}\right| \\
\leq \;& \varepsilon\left((k_0+1)g(Y_{(n-k,n)})\Big(\frac{Y_{(n-k_0,n)}}{Y_{(n-k,n)}}\Big)^{\varepsilon} + \sum_{i=k_0+2}^{k} g(Y_{(n-k,n)})\Big(\frac{Y_{(n-i+1,n)}}{Y_{(n-k,n)}}\Big)^{\varepsilon}\right)
\end{aligned} \tag{C.37}$$

Since $Y_{(n-i+1,n)} \geq Y_{(n-k_0,n)}$ for $i = 1, \cdots, k_0+1$, we further obtain

$$\max_{0\leq k_0\leq k}\Big(\frac{(k-k_0)}{kg(Y_{(n-k,n)})}|R_{k_0,k} - S_{k_0,k}|\Big) \leq \frac{\varepsilon}{k}\sum_{i=1}^{k}\Big(\frac{Y_{(n-i+1,n)}}{Y_{(n-k,n)}}\Big)^{\varepsilon} \tag{C.38}$$

over the event $\{Y_{(n-k,n)} > t_\varepsilon\}$. The upper bound in (C.38) can be bounded by 2ε over the event $\{(1/k)\sum_{i=1}^{k}(Y_{(n-i+1,n)}/Y_{(n-k,n)})^{\varepsilon} < 2\}$.

We have already proved that $\mathbb{P}(Y_{(n-k,n)} > t_\varepsilon) \to 1$. Thus, to complete the proof of Relation (C.35), it only remains to show that

$$\mathbb{P}\left(\Big\{\frac{1}{k}\sum_{i=1}^{k}(\frac{Y_{(n-i+1,n)}}{Y_{(n-k,n)}})^{\varepsilon} < 2\Big\}\right) \to 1. \tag{C.39}$$

In this direction, from Relation (C.10), we observe that

$$\frac{1}{k}\sum_{i=1}^{k}\Big(\frac{Y_{(n-i+1,n)}}{Y_{(n-k,n)}}\Big)^{\varepsilon} \stackrel{d}{=} \frac{1}{k}\sum_{i=1}^{k}\Big(\frac{\Gamma_{i+1}}{\Gamma_{k+1}}\Big)^{-\varepsilon} = \frac{1}{k}\sum_{i=1}^{k}U_{i,k}^{-\varepsilon} \stackrel{\mathbb{P}}{\longrightarrow} \frac{1}{1-\varepsilon}$$

where the last convergence follows from weak law of large numbers. Thus, Relation (C.39) holds as long as $\varepsilon < 0.5$.

This completes the proof for $\rho = 0$. □

Lemma C.8. *Suppose g is ρ-varying for any $\rho \in (-\infty, \infty)$ and $Y_{(n-k,n)}$ is the $(k+1)^{th}$ order statistic for n observations from* Pareto$(1,1)$*, then*

$$\frac{g(Y_{(n-k,n)})}{g(n/k)} \xrightarrow{\mathbb{P}} 1 \tag{C.40}$$

provided $k \to \infty$, $n \to \infty$ and $k/n \to \infty$.

Proof. Since g is ρ varying, g may be expressed as $g(t) = t^\rho \ell(t)$, for some slowly varying function $\ell(\cdot)$. Thus, we have

$$\frac{g(Y_{(n-k,n)})}{g(n/k)} = \Big(\frac{Y_{(n-k,n)}}{n/k}\Big)^\rho \frac{\ell(Y_{(n-k,n)})}{\ell(n/k)}$$

From Relation (C.10), $Y_{(n-k,n)} \stackrel{d}{=} \Gamma_{n+1}/\Gamma_{k+1}$ and therefore, by WLLN, we have $Y_{(n-k,n)}/(n/k) \xrightarrow{\mathbb{P}} 1$. Since x^ρ is a continuous mapping, therefore,

$(Y_{(n-k,n)}/(n/k))^\rho \xrightarrow{\mathbb{P}} 1$.

Thus to prove Relation (C.40), it suffices to show $\ell(Y_{(n-k,n)})/\ell(n/k) \xrightarrow{\mathbb{P}} 1$. In this direction, observe that for any $\delta > 0$, we have

$$\begin{aligned}
& \mathbb{P}\Big(\Big|\frac{\ell(Y_{(n-k,n)})}{\ell(n/k)} - 1\Big| > \varepsilon\Big) \\
\leq\ & \mathbb{P}\Big(\Big|\frac{\ell(Y_{(n-k,n)})}{\ell(n/k)} - 1\Big| > \varepsilon, \Big|\frac{Y_{(n-k,n)}}{n/k} - 1\Big| \leq \delta\Big) + \mathbb{P}\Big(\Big|\frac{Y_{(n-k,n)}}{n/k} - 1\Big| > \delta\Big) \\
\leq\ & \mathbb{P}\Big(\sup_{\lambda \in [1-\delta, 1+\delta)} \Big|\frac{\ell(\lambda n/k)}{\ell(n/k)} - 1\Big| > \varepsilon\Big) + \mathbb{P}\Big(\Big|\frac{Y_{(n-k,n)}}{n/k} - 1\Big| > \delta\Big)
\end{aligned}$$

For δ small enough, the first term on the right hand side goes to 0 by Theorem 1.5.2 on page 22 in [10]. Also, for δ small enough, the second term goes to 0 since $Y_{(n-k,n)}/(n/k) \xrightarrow{\mathbb{P}} 1$.

This completes the proof. □

Lemma C.9.

$$\max_{0 \leq k_0 < k} \Big|\frac{k-k_0}{k g(Y_{(n-k,n)})} S_{k_0,k} + \frac{c}{1+\rho}\Big(\frac{k_0}{k}\Big)^{1+\rho} - \frac{c}{1+\rho}\Big| \xrightarrow{\mathbb{P}} 0. \tag{C.41}$$

where $S_{k_0,k}$ is defined in Relation (C.32).

Proof. The proof of Relation (C.41) involves two cases: $\rho > 0$ and $\rho = 0$.

Case $\rho > 0$: Using the expression of $S_{k_0,k}$ in Relation (C.32), we get

$$\begin{aligned}
& \frac{k-k_0}{k g(Y_{(n-k,n)})} S_{k_0,k} \\
=\ & -\frac{c}{k\rho}\Bigg((k_0+1)\Big(\frac{Y_{(n-k_0,n)}}{Y_{(n-k,n)}}\Big)^{-\rho} + \sum_{i=k_0+2}^{k}\Big(\frac{Y_{(n-i+1,n)}}{Y_{(n-k,n)}}\Big)^{-\rho} - k\Bigg)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{c}{k\rho}\sum_{i=1}^{k_0}\left\{\left(\frac{Y_{(n-i+1,n)}}{Y_{(n-k,n)}}\right)^{-\rho}-\left(\frac{Y_{(n-k_0,n)}}{Y_{(n-k,n)}}\right)^{-\rho}\right\}\\
&- \frac{c}{k\rho}\sum_{i=1}^{k}\left\{\left(\frac{Y_{(n-i+1,n)}}{Y_{(n-k,n)}}\right)^{-\rho}-1\right\} \qquad \text{(C.42)}
\end{aligned}
$$

Expressing the order statistics of Pareto in terms of Gamma random variables as in Relation (C.10), we get

$$
\begin{aligned}
&\frac{k-k_0}{kg(Y_{(n-k,n)})}S_{k_0,k}+\frac{c}{1+\rho}\left(\frac{k_0}{k}\right)^{1+\rho}-\frac{c}{1+\rho}\\
\stackrel{d}{=}\ &\underbrace{\frac{c}{1+\rho}\left(\frac{k_0}{k}\right)^{1+\rho}+\frac{c}{k\rho}\sum_{i=1}^{k_0}\left\{\left(\frac{\Gamma_{i+1}}{\Gamma_{k+1}}\right)^{\rho}-\left(\frac{\Gamma_{k_0+1}}{\Gamma_{k+1}}\right)^{\rho}\right\}}_{B_{k_0,k}}\\
-\ &\underbrace{\left(\frac{c}{k\rho}\sum_{i=1}^{k}\left\{\left(\frac{\Gamma_{i+1}}{\Gamma_{k+1}}\right)^{\rho}-1\right\}+\frac{c}{1+\rho}\right)}_{A_k}
\end{aligned}
$$

In view of the above result, to prove (C.41), we first show that $\max_{0\le k_0<k}|A_k|\xrightarrow{a.s.}0$.

Note that, by Relation (C.7), we have $|(1/k)\sum_{i=1}^{k}(\Gamma_{i+1}/\Gamma_{k+1})^{\rho}-1/(1+\rho)|\xrightarrow{a.s.}0$. This implies that there exists Ω with $\mathbb{P}(\Omega)=1$ such that for any $\omega\in\Omega$,

$$
\begin{aligned}
\left|A_k(\omega)\right| &= \left|\frac{c}{\rho k}\sum_{i=1}^{k}\left(\frac{\Gamma_{i+1}}{\Gamma_{k+1}}\right)^{\rho}(\omega)-\frac{c}{\rho}+\frac{c}{1+\rho}\right|\\
&= \left|\frac{c}{k\rho}\sum_{i=1}^{k}\left(\frac{\Gamma_{i+1}}{\Gamma_{k+1}}\right)^{\rho}(\omega)-\frac{c}{\rho(1+\rho)}\right|\to 0
\end{aligned}
$$

We next show that $\max_{0\le k_0<k}B_{k_0,k}\xrightarrow{a.s.}0$. For this observe that for any $\omega\in\Omega$,

$$
\begin{aligned}
&\max_{0\le k_0<M}B_{k_0,k}(\omega)\\
\le\ &\max_{0\le k_0<M}\left\{\frac{c}{1+\rho}\left(\frac{k_0}{k}\right)^{1+\rho}+\frac{c}{k\rho}\sum_{i=1}^{k_0}\left|\left(\frac{\Gamma_{i+1}}{\Gamma_{k+1}}\right)^{\rho}(\omega)-\left(\frac{\Gamma_{k_0+1}}{\Gamma_{k+1}}\right)^{\rho}(\omega)\right|\right\}\\
\le\ &\max_{0\le k_0<M}\left\{\frac{c}{1+\rho}\left(\frac{k_0}{k}\right)^{1+\rho}+\frac{2ck_0}{k\rho}\right\}\\
&(\text{since } (\Gamma_i/\Gamma_{k+1})^{\rho}\le 1,\ 1\le i\le k,\ \rho>0)\\
\le\ &\frac{cM^{1+\rho}/(1+\rho)+2cM/\rho}{k}=\frac{B_{0M}}{k}. \qquad \text{(C.43)}
\end{aligned}
$$

Additionally, we have

$$\begin{aligned}
&\max_{M\le k_0<k} B_{k_0,k}(\omega)\\
\le\ &\max_{M\le k_0<k}\left|\frac{c}{1+\rho}\left(\frac{k_0}{k}\right)^{1+\rho}+\frac{c}{k\rho}\sum_{i=1}^{k_0}\left\{\left(\frac{\Gamma_{i+1}}{\Gamma_{k+1}}\right)^{\rho}-\left(\frac{\Gamma_{k_0+1}}{\Gamma_{k+1}}\right)^{\rho}\right\}(\omega)\right|\\
\le\ &\max_{M\le k_0<k} c\left(\frac{k_0}{k}\right)^{1+\rho}\left|\frac{\rho}{1+\rho}+\left(\frac{k_0}{k}\right)^{\rho}\frac{1}{k_0}\sum_{i=1}^{k_0}\left\{\left(\frac{\Gamma_{i+1}}{\Gamma_{k+1}}\right)^{\rho}-\left(\frac{\Gamma_{k_0+1}}{\Gamma_{k+1}}\right)^{\rho}\right\}(\omega)\right|\\
\le\ &c\max_{M\le k_0<k}\left|\frac{\rho}{1+\rho}+\left(\frac{\Gamma_{k_0+1}/k_0}{\Gamma_{k+1}/k}\right)^{\rho}(\omega)\underbrace{\left\{\frac{1}{k_0}\sum_{i=1}^{k_0}\left(\frac{\Gamma_{i+1}}{\Gamma_{k_0+1}}\right)^{\rho}(\omega)-1\right\}}_{C_{k_0}(\omega)}\right|\\
=\ &c\max_{M\le k_0<k}\left|\left\{\left(\frac{\Gamma_{k_0+1}/k_0}{\Gamma_{k+1}/k}\right)^{\rho}-1\right\}(C_{k_0}(\omega)-1)+(C_{k_0}(\omega)-1)+\frac{\rho}{1+\rho}\right|.
\end{aligned}\tag{C.44}$$

Since $\Gamma_{i+1}<\Gamma_{k_0+1}$ and $\rho>0$, thereby $|C_{k_0}|<1$. This allows us to simplify the upper bound in Relation (C.44) as

$$\begin{aligned}
&\max_{M\le k_0<k} B_{k_0,k}(\omega)\\
\le\ &\underbrace{c\sup_{M\le k_0}\left|C_{k_0}(\omega)-\frac{1}{1+\rho}\right|}_{B_{1M}(\omega)}+2c\underbrace{\sup_{M\le k_0,k}\left|\left(\frac{\Gamma_{k_0+1}/k_0}{\Gamma_{k+1}/k}\right)^{\rho}(\omega)-1\right|}_{B_{2M}(\omega)}
\end{aligned}$$

Thus, we obtain

$$\max_{0\le k_0<k} B_{k_0,k}(\omega)\le\frac{B_{0M}}{k}+cB_{1M}(\omega)+cB_{2M}(\omega).$$

Taking lim sup w.r.t to k on both sides, we get

$$\limsup_{k\to\infty}\max_{0\le k_0<k} B_{k_0,k}(\omega)\le c(B_{1M}(\omega)+B_{2M})(\omega).\tag{C.45}$$

By using Lemmas C.1 and C.3, we can show that there exist $\tilde{\Omega}$ with $\mathbb{P}(\tilde{\Omega})=1$ such that for all $\omega\in\tilde{\Omega}$, $B_{1M}(\omega)\to 0$ and $B_{2M}(\omega)\to 0$.

Thus, taking lim sup w.r.t M on both sides of Relation (C.45), we get $\max_{0\le k_0<k} B_{k_0,k}\xrightarrow{a.s.}0$. This completes the proof for $\rho>0$.

Case $\rho=0$*:* Using the expression of $S_{k_0,k}$ in Relation (C.32), we get

$$\begin{aligned}
&\frac{k-k_0}{kg(Y_{(n-k,n)})}S_{k_0,k}+\frac{ck_0}{k}-c\\
=\ &\frac{c}{k}\left((k_0+1)\log\left(\frac{Y_{(n-k_0,n)}}{Y_{(n-k,n)}}\right)+\sum_{i=k_0+2}^{k}\log\left(\frac{Y_{(n-i+1,n)}}{Y_{(n-k,n)}}\right)\right)-\frac{c(k-k_0)}{k}
\end{aligned}$$

$$\stackrel{d}{=} \frac{c(k-k_0)}{k}\widehat{\xi}^{**}_{k_0,k} - \frac{c(k-k_0)}{k} \stackrel{d}{=} c\Big(\frac{\Gamma_{k-k_0}}{k} - \frac{k-k_0}{k}\Big), \tag{C.46}$$

where $\widehat{\xi}^{**}_{k_0,k}$ is the trimmed Hill estimator in Relation (2.2) with X_i's replaced by the i.i.d. Pareto$(1,1)$. The last distribution equality in Relation (C.46) follows from Relation (2.4).

Thus, to prove Relation (C.41), we shall next show $\max_{0\le k_0<k} |\Gamma_{k-k_0} - (k-k_0)|/k \xrightarrow{a.s.} 0$. In this direction, for any $\omega \in \Omega$, we have

$$\begin{aligned}
&\max_{0\le k_0<k} \frac{|\Gamma_{k-k_0}(\omega) - (k-k_0)|}{k} \\
= &\max_{0\le k_0<k} \frac{(k-k_0)}{k}\Big|\frac{\Gamma_{k-k_0}}{k-k_0}(\omega) - 1\Big| \\
\le &\frac{M}{k}\max_{0\le k-k_0<M}\Big|\frac{\Gamma_{k-k_0}}{k-k_0}(\omega) - 1\Big| + \sup_{k-k_0\ge M}\Big|\frac{\Gamma_{k-k_0}}{k-k_0}(\omega) - 1\Big| \\
\le &\frac{M}{k}\underbrace{\sup_n\Big|\frac{\Gamma_n}{n}(\omega) - 1\Big|}_{B_0(\omega)} + \underbrace{\sup_{n\ge M}\Big|\frac{\Gamma_n}{n}(\omega) - 1\Big|}_{B_{1M}(\omega)}
\end{aligned} \tag{C.47}$$

By SLLN, there exists Ω with $\mathbb{P}(\Omega) = 1$ such that for every $\omega \in \Omega$, $|\Gamma_n(\omega)/n - 1| \to 0$ as $n \to \infty$. This implies that $B_0(\omega)$ is bounded and also that $B_{1M}(\omega) \to 0$ as $M \to \infty$.

Thus, first taking lim sup with respect to k followed by lim sup with respect to M on both sides of Relation (C.47), the proof follows. □

C.3.3. Consistency of the weighted sequential testing

Proof of Theorem 3.9. From Relation (2.6), we have

$$\begin{aligned}
k^\delta \max_{0\le k_0<h(k)} |T_{k_0,k} - T^*_{k_0,k}| &= k^\delta \max_{0\le k_0<h(k)} \frac{k-k_0-1}{k-k_0}\Big|\frac{\widehat{\xi}_{k_0+1,k}}{\widehat{\xi}_{k_0,k}} - \frac{\widehat{\xi}^*_{k_0+1,k}}{\widehat{\xi}^*_{k_0,k}}\Big| \\
&\le k^\delta \max_{0\le k_0<h(k)} \underbrace{\Big|\frac{\widehat{\xi}_{k_0+1,k}}{\widehat{\xi}_{k_0,k}} - \frac{\widehat{\xi}^*_{k_0+1,k}}{\widehat{\xi}^*_{k_0,k}}\Big|}_{W_{k_0,k}}
\end{aligned}$$

where the last inequality holds since $k-k_0-1 \le k-k_0$ for $0 \le k_0 < h(k)$. We complete the proof by showing that $k^\delta \max_{0\le k_0<h(k)} W_{k_0,k} \xrightarrow{\mathbb{P}} 0$. To this end, we observe that

$$\begin{aligned}
&W_{k_0,k} \\
\le &\Big|\frac{\widehat{\xi}_{k_0+1,k}}{\widehat{\xi}_{k_0,k}} - \frac{\widehat{\xi}^*_{k_0+1,k}}{\widehat{\xi}_{k_0,k}} - \frac{cAk^{-\delta}}{(1+\rho)\widehat{\xi}_{k_0,k}}\Big| + \Big|\frac{cAk^{-\delta}}{(1+\rho)\widehat{\xi}_{k_0,k}} - \frac{cAk^{-\delta}}{(1+\rho)\widehat{\xi}_{k_0,k}}\frac{\widehat{\xi}^*_{k_0+1,k}}{\widehat{\xi}^*_{k_0,k}}\Big|
\end{aligned}$$

$$
\begin{aligned}
&+ \left| \frac{\widehat{\xi}^*_{k_0+1,k}}{\widehat{\xi}_{k_0,k}} - \frac{\widehat{\xi}^*_{k_0+1,k}}{\widehat{\xi}^*_{k_0,k}} + \frac{cAk^{-\delta}}{(1+\rho)\widehat{\xi}_{k_0,k}} \frac{\widehat{\xi}^*_{k_0+1,k}}{\widehat{\xi}^*_{k_0,k}} \right| \\
&= \frac{\left(\left| R_{k_0+1,k} - \frac{cAk^{-\delta}}{1+\rho} \right| + \frac{|c|Ak^{-\delta}}{(1+\rho)} \left| 1 - \frac{\widehat{\xi}^*_{k_0+1,k}}{\widehat{\xi}^*_{k_0,k}} \right| + \frac{\widehat{\xi}^*_{k_0+1,k}}{\widehat{\xi}^*_{k_0,k}} \left| \frac{cAk^{-\delta}}{(1+\rho)} - R_{k_0,k} \right| \right)}{\widehat{\xi}_{k_0,k}}
\end{aligned}
$$

where $R_{k_0,k}$ is defined in Relation (3.5). Thus, with

$$
M_{1k} := k^{\delta} \max_{0 \le k_0 < h(k)} \left| R_{k_0,k} - \frac{cAk^{-\delta}}{1+\rho} \right| \quad \text{and} \quad B_{k_0,k} := \frac{\widehat{\xi}^*_{k_0+1,k}}{\widehat{\xi}^*_{k_0,k}},
$$

the quantity $k^{\delta} \max_{0 \le k_0 < h(k)} W_{k_0,k}$ is bounded above as follows

$$
\begin{aligned}
&\max_{0 \le k_0 < h(k)} k^{\delta} W_{k_0,k} \\
&\le \underbrace{\left(M_{1k} \max_{0 \le k_0 \le h(k)} (1 + B_{k_0,k}) + \frac{|c|A}{(1+\rho)} \max_{0 \le k_0 \le h(k)} |1 - B_{k_0,k}| \right)}_{\Delta_k} \max_{0 \le k_0 < h(k)} \frac{1}{\widehat{\xi}_{k_0,k}}
\end{aligned} \tag{C.48}
$$

Theorem 3.5 implies $M_{1k} \xrightarrow{\mathbb{P}} 0$ as $k \to \infty$. On the other hand, by Relation (C.19),

$$
\begin{aligned}
\max_{0 \le k_0 \le h(k)} |1 - B_{k_0,k}| &\overset{d}{=} \max_{0 \le k_0 \le h(k)} \left| 1 - \frac{\Gamma_{k-k_0-1}/(k-k_0-1)}{\Gamma_{k-k_0}/(k-k_0)} \right| \\
&\le \frac{1}{1 - h(k)/k} \max_{k-h(k) \le i \le k} \left| \frac{\Gamma_i / i}{\Gamma_{i+1}/(i+1)} - 1 \right| \xrightarrow{a.s.} 0,
\end{aligned}
$$

where the last convergence is a consequence of Relation (C.6). The fact that $\max_{0 \le k_0 \le h(k)} (1 + B_{k_0,k}) \xrightarrow{a.s.} 0$ also implies $\max_{0 \le k_0 \le h(k)} (1 + B_{k_0,k}) = O_{\mathbb{P}}(1)$. Thus Δ_k in Relation (C.48) converges to 0.

We shall end the proof by showing that $\min_{0 \le k_0 < h(k)} |\widehat{\xi}_{k_0,k}|$ is bounded away from 0 in probability which in view of Relation (C.48), implies the convergence of $\max_{0 \le k_0 < h(k)} k^{\delta} W_{k_0,k}$ to 0 in probability. To this end, we have,

$$
\min_{0 \le k_0 < h(k)} \widehat{\xi}_{k_0,k} \ge \min_{0 \le k_0 < h(k)} \widehat{\xi}^*_{k_0,k} - \max_{0 \le k_0 < h(k)} |\widehat{\xi}_{k_0,k} - \widehat{\xi}^*_{k_0,k}| \tag{C.49}
$$

For $\delta > 0$, Theorem 3.5 implies $\max_{0 \le k_0 < h(k)} |\widehat{\xi}_{k_0,k} - \widehat{\xi}^*_{k_0,k}| \xrightarrow{\mathbb{P}} 0$. Therefore $\min_{0 \le k_0 < h(k)} \widehat{\xi}_{k_0,k}$ is bounded away from 0 as long as $\min_{0 \le k_0 < h(k)} \widehat{\xi}^*_{k_0,k}$ is bounded away from 0. This is easy to show because

$$
\min_{0 \le k_0 < h(k)} \widehat{\xi}^*_{k_0,k} \overset{d}{=} \min_{0 \le k_0 < h(k)} \frac{\Gamma_{k-k_0}}{k-k_0} \ge 1 - \max_{k-h(k) \le i < k} \left| \frac{\Gamma_i}{i} - 1 \right| \xrightarrow{a.s.} 1
$$

where the last equality in distribution is due to Relation (2.4) and the last convergence is a direct consequence of Relation (C.5). This completes the proof. □

Proof of Theorem 3.11. *Proof of Relation* (3.15): By Relation (2.7), we have

$$\begin{aligned} & k^{(\delta^*-1)} \max_{0\le k_0<h(k)} |U_{k_0,k} - U^*_{k_0,k}| \\ = \; & 2k^{(\delta^*-1)} \max_{0\le k_0<h(k)} \Big| |(T_{k_0,k})^{k-k_0-1} - 0.5| - |(T^*_{k_0,k})^{k-k_0-1} - 0.5| \Big| \\ \le \; & 2k^{(\delta^*-1)} \max_{0\le k_0<h(k)} \Big| (T_{k_0,k})^{k-k_0-1} - (T^*_{k_0,k})^{k-k_0-1} \Big| \\ \le \; & 2k^{(\delta^*-1)} \max_{0\le k_0<h(k)} \Big| \Big(\frac{T_{k_0,k}}{T^*_{k_0,k}}\Big)^{k-k_0-1} - 1 \Big| \end{aligned} \tag{C.50}$$

where the last bound follows $T^*_{k_0,k} \le 1$ (see Proposition 2.7). In view of Relation (C.50), to prove Relation (3.15), it suffices to show

$$k^{(\delta^*-1)} \max_{0\le k_0<h(k)} \Big| \Big(\frac{T_{k_0,k}}{T^*_{k_0,k}}\Big)^{k-k_0-1} - 1 \Big| \xrightarrow{\mathbb{P}} 0. \tag{C.51}$$

To this end, we begin by showing

$$k^{\delta^*} \max_{0\le k_0<h(k)} \Big| \frac{T_{k_0,k}}{T^*_{k_0,k}} - 1 \Big| \xrightarrow{\mathbb{P}} 0. \tag{C.52}$$

In this direction, observe that

$$k^{\delta^*} \max_{0\le k_0<h(k)} \Big| \frac{T_{k_0,k}}{T^*_{k_0,k}} - 1 \Big| \le \frac{1}{\min_{0\le k_0<h(k)} T^*_{k_0,k}} \max_{0\le k_0<h(k)} \underbrace{k^{\delta^*} |T_{k_0,k} - T^*_{k_0,k}|}_{\Delta_k},$$

where $\Delta_k \xrightarrow{\mathbb{P}} 0$ by Relation (3.14). Thus, Relation (C.52) holds as long as $\min_{0\le k_0<h(k)} T^*_{k_0,k}$ is bounded away from 0 in probability as shown next.

$$\begin{aligned} \min_{0\le k_0<h(k)} T^*_{k_0,k} & \overset{d}{=} \min_{0\le k_0<h(k)} \frac{\Gamma_{k-k_0-1}/(k-k_0-1)}{\Gamma_{k-k_0}/(k-k_0)} \\ & \ge 1 - \max_{k-h(k)\le i<k} \Big| \frac{\Gamma_i/i}{\Gamma_{i+1}/(i+1)} - 1 \Big| \xrightarrow{a.s.} 1, \end{aligned}$$

where the last convergence is a direct consequence of Relation (C.6).

Finally to prove Relation (C.51), we shall equivalently show that for every subsequence $\{k_l\}$, there exists a further subsequence $\widetilde{k}$ such that

$$\widetilde{k}^{(\delta^*-1)} \max_{0\le k_0<h(\widetilde{k})} \Big| \Big(\frac{T_{k_0,\widetilde{k}}}{T^*_{k_0,\widetilde{k}}}\Big)^{\widetilde{k}-k_0-1} - 1 \Big| \xrightarrow{a.s.} 0. \tag{C.53}$$

This is shown next as follows. In view of Relation (C.52), for every subsequence $\{k_l\}$, there exists a further subsequence $\widetilde{k}$ such that

$$\widetilde{k}^{\delta^*} \max_{0\le k_0<h(\widetilde{k})} \Big|\frac{T_{k_0,\widetilde{k}}}{T^*_{k_0,\widetilde{k}}} - 1\Big| \xrightarrow{a.s.} 0.$$

Hence, there is event an Ω with $\mathbb{P}(\Omega) = 1$ such that for every $\epsilon > 0$, there exist a $M = M(\omega, \epsilon)$

$$1 - \frac{\epsilon}{\widetilde{k}^{\delta^*}} \le \Big(\frac{T_{k_0,\widetilde{k}}}{T^*_{k_0,\widetilde{k}}}\Big)(\omega) \le 1 + \frac{\epsilon}{\widetilde{k}^{\delta^*}}, \quad \text{for all } \widetilde{k} \ge M,\ 0 \le k_0 < h(\widetilde{k}) \text{ and } \omega \in \Omega \tag{C.54}$$

Therefore,

$$\begin{aligned} & \underbrace{\widetilde{k}^{(\delta^*-1)}\Big(\Big(1 - \frac{\epsilon}{\widetilde{k}^{\delta^*}}\Big)^{\widetilde{k}-h(\widetilde{k})-1} - 1\Big)}_{-a_{\widetilde{k}}} \\ \le\ & \widetilde{k}^{(\delta^*-1)}\Big(\Big(\frac{T_{k_0,\widetilde{k}}}{T^*_{k_0,\widetilde{k}}}\Big)^{\widetilde{k}-k_0-1}(\omega) - 1\Big) \\ \le\ & \underbrace{\widetilde{k}^{(\delta^*-1)}\Big(\Big(1 + \frac{\epsilon}{\widetilde{k}^{\delta^*}}\Big)^{\widetilde{k}-1} - 1\Big)}_{b_{\widetilde{k}}} \end{aligned}$$

which equivalently implies

$$\widetilde{k}^{(\delta^*-1)} \max_{0\le k_0<h(\widetilde{k})} \Big|\Big(\frac{T_{k_0,\widetilde{k}}}{T^*_{k_0,\widetilde{k}}}\Big)^{\widetilde{k}-k_0-1}(w) - 1\Big| \le a_{\widetilde{k}} \vee b_{\widetilde{k}} \tag{C.55}$$

Note that both the sequences $a_{\widetilde{k}}$ and $b_{\widetilde{k}}$ converge to ϵ as $\widetilde{k} \to \infty$. Thereby, taking limsup w.r.t $\widetilde{k}$ on both sides of Relation (C.55), we get

$$\limsup_{\widetilde{k}\to\infty} \widetilde{k}^{(\delta^*-1)} \max_{0\le k_0<h(\widetilde{k})} \Big|\Big(\frac{T_{k_0,\widetilde{k}}}{T^*_{k_0,\widetilde{k}}}\Big)^{\widetilde{k}-k_0-1}(w) - 1\Big| \le \epsilon \tag{C.56}$$

Since Relation (C.56) holds for all $\epsilon > 0$ and $\omega \in \Omega$ with $\mathbb{P}(\Omega) = 1$, we have

$$\widetilde{k}^{(\delta^*-1)} \max_{0\le k_0<h(\widetilde{k})} \Big|\Big(\frac{T_{k_0,\widetilde{k}}}{T^*_{k_0,\widetilde{k}}}\Big)^{\widetilde{k}-k_0-1} - 1\Big| \xrightarrow{a.s.} 0$$

This entails the proof of the convergence in probability of Relation (C.51).

Proof of Relation (3.16). To this end, we show that $\mathbb{P}_{\mathcal{H}_0}(\widehat{k}_0 = 0) \to 1 - q$.

We first show that $\limsup \mathbb{P}_{\mathcal{H}_0}(\widehat{k}_0 = 0) \leq 1 - q$ as follows.

$$\begin{aligned}
\mathbb{P}_{\mathcal{H}_0}(\widehat{k}_0 = 0) &= \mathbb{P}_{\mathcal{H}_0}\Big(\underbrace{\bigcap_{i=0}^{h(k)} \{U_{i,k} < (1-q)^{ca^{k-i-1}}\}}_{A_k} \Big) \\
&\leq \mathbb{P}_{\mathcal{H}_0}\Big(A_k \cap \underbrace{\{k^{(\delta^*-1)} \max_{0 \leq i \leq h(k)} (U_{i,k} - U^*_{i,k}) < \epsilon\}}_{B_{1k}} \Big) \\
&+ \mathbb{P}_{\mathcal{H}_0}(\{k^{(\delta^*-1)} \max_{0 \leq i \leq h(k)} (U_{i,k} - U^*_{i,k}) > \epsilon\}) \\
&\leq \mathbb{P}_{\mathcal{H}_0}\Big(\underbrace{\bigcap_{i=0}^{h(k)} \{U^*_{i,k} < (1-q)^{ca^{k-i-1}} + \epsilon k^{(1-\delta^*)}\}}_{A^*_{1k}} \Big) + \mathbb{P}_{\mathcal{H}_0}(B^c_{1k}) \\
&\quad (\text{since } A_k \cap B_{1k} \implies A^*_{1k})
\end{aligned}$$

By Relation (3.15), $\mathbb{P}_{\mathcal{H}_0}(B^c_{1k}) \to 0$. Additionally, we have shown below that $\limsup_{k\to\infty} \mathbb{P}_{\mathcal{H}_0}(A^*_{1k}) \leq 1 - q$ which implies

$$\limsup \mathbb{P}_{\mathcal{H}_0}(\widehat{k}_0 = 0) = \limsup_{k\to\infty} \mathbb{P}(A_k) \leq 1 - q$$

.

Since $U^*_{i,k}$ are i.i.d. $U(0,1)$, therefore

$$\begin{aligned}
\mathbb{P}_{\mathcal{H}_0}(A^*_{1k}) &= (1-q)^{\sum_{i=0}^{h(k)} ca^{k-i-1}} \prod_{i=0}^{h(k)} \Big(1 + \frac{\epsilon k^{(1-\delta^*)}}{(1-q)^{ca^{k-i-1}}}\Big) \\
&\leq \underbrace{(1-q)^{\sum_{i=0}^{h(k)} ca^{k-i-1}}}_{c_{0k}} \underbrace{\Big(1 + \frac{\epsilon}{(1-q)k^{(\delta^*-1)}}\Big)^{h(k)}}_{c_{1k}} \\
&\quad (\text{since } (1-q)^{ca^{k-i-1}} \geq (1-q) \ ).
\end{aligned}$$

Since $h(k) = o(k)$, it is easy to show that $\limsup_{k\to\infty} c_{0k} = \limsup_{k\to\infty} (1-q)^{\sum_{i=0}^{h(k)} ca^{k-i-1}} = 1 - q$. For $\delta^* \geq 2$, $h(k) = o(k^{(\delta^*-1)})$ which implies $\limsup_{k\to\infty} c_{1k} \leq 1$. Thus,

$$\limsup_{k\to\infty} \mathbb{P}_{\mathcal{H}_0}(A^*_{1k}) \leq (1-q)$$

.

Finally, we show that $\liminf \mathbb{P}_{\mathcal{H}_0}(\widehat{k}_0 = 0) = \liminf_{k\to\infty} \mathbb{P}_{\mathcal{H}_0}(A_k) \geq 1 - q$ as follows:

$$\mathbb{P}_{\mathcal{H}_0}(A_k) \geq \mathbb{P}_{\mathcal{H}_0}\Big(A_k \cap \underbrace{\{k^{(\delta^*-1)} \max_{0 \leq i < k} (U_{i,k} - U^*_{i,k}) > -\epsilon\}}_{B_{2k}} \Big)$$

$$\begin{aligned} &\geq \mathbb{P}_{\mathcal{H}_0}\Big(\underbrace{\bigcap_{i=0}^{h(k)} \{U^*_{i,k} < (1-q)^{ca^{k-i-1}} - \epsilon k^{(1-\delta^*)}\}}_{A^*_{2k}} \cap B_{2k} \Big) \\ &= \mathbb{P}_{\mathcal{H}_0}(A^*_{2k}) - \mathbb{P}_{\mathcal{H}_0}(B^c_{2k}), \end{aligned}$$

since $A^*_{2k} \cap B_{2k} \implies A_k \cap B_{2k}$. By Relation (3.15), $\mathbb{P}(B^c_{2k}) \to 0$. Additionally, it has been shown below that $\liminf_{k\to\infty} \mathbb{P}_{\mathcal{H}_0}(A^*_{2k}) \geq 1-q$ which implies $\liminf_{k\to\infty} \mathbb{P}(A_k) \geq 1-q$.

$$\begin{aligned} \mathbb{P}_{\mathcal{H}_0}(A^*_{2k}) &= (1-q)^{\sum_{i=0}^{h(k)} ca^{k-i-1}} \prod_{i=0}^{h(k)} \Big(1 - \frac{\epsilon k^{(1-\delta^*)}}{(1-q)^{ca^{k-i-1}}}\Big) \\ &\geq \underbrace{(1-q)^{\sum_{i=0}^{h(k)} ca^{k-i-1}}}_{c_{0k}} \underbrace{\Big(1 - \frac{\epsilon}{(1-q)k^{(\delta^*-1)}}\Big)^{h(k)}}_{c_{2k}} \\ &\qquad (\text{since } (1-q)^{ca^{k-i-1}} \geq (1-q)) \end{aligned}$$

Since $h(k) = o(k)$, it is easy to show that $\liminf_{k\to\infty} c_{0k} = \liminf_{k\to\infty} (1-q)^{\sum_{i=0}^{h(k)} ca^{k-i-1}} = 1-q$. For $\delta^* \geq 2$, $h(k) = o(k^{(\delta^*-1)})$ which implies $\liminf_{k\to\infty} c_{2k} \geq 1$. Thus,

$$\liminf_{k\to\infty} \mathbb{P}_{\mathcal{H}_0}(A^*_{2k}) \geq (1-q)$$

This completes the proof. □

Acknowledgements

We thank two anonymous referees and the editor of Electronic Journal of Statistics for insightful remarks, which helped us improve the paper. Specifically, the discussion in Section A.1 is entirely motivated by a point raised by a referee.

References

[1] M. Kallitsis, S. A. Stoev, S. Bhattacharya, and G. Michailidis. Amon: An open source architecture for online monitoring, statistical analysis, and forensics of multi-gigabit streams. *IEEE Journal on Selected Areas in Communications*, 34(6):1834–1848, June 2016.

[2] I. B. Aban, M. M. Meerschaert, and A. K. Panorska. Parameter Estimation for the Truncated Pareto Distribution. *Journal of the American Statistical Association*, 101:270–277, 2006. MR2268044

[3] M. Ahsanullah, V. Nevzorov, and M. Shakil. An introduction to order statistics, volume 3 of *Atlantis Studies in Probability and Statistics*. Atlantis Press, Paris, 2013. MR3025012

[4] J. Beirlant, Ch. Bouquiaux, and B. Werker. Semiparametric lower bounds for tail index estimation. *Journal of Statistical Planning and Inference*, 136(3):705–729, 2006. MR2181974

[5] J. Beirlant, P. Vynckier and J. L. Teugels. Tail Index Estimation, Pareto Quantile Plots, and Regression Diagnostics. *Journal of the American Statistical Association*, 436(91):1659–1667, 1996. MR1439107

[6] J. Beirlant, I. Fraga Alves and I. Gomes. Tail fitting for truncated and non-truncated Pareto-type distributions. *Extremes*, 19(3):429–462, 2016. MR3535961

[7] J. Beirlant, Y. Goegebeur, J. Teugels, and J. Segers. Statistics of extremes: Theory and applications. *Wiley Series in Probability and Statistics*. John Wiley & Sons, Ltd., Chichester, 2004. MR2108013

[8] J. Beirlant, A. Guillou, G. Dierckx, and A. Fils-Villetard. Estimation of the extreme value index and extreme quantiles under random censoring. *Extremes*, 10(3):151–174, 2007. MR2415636

[9] Package CASdatasets. *freclaimset*, http://cas.uqam.ca/pub/R/web/CASdatasets-manual.pdf p. 42.

[10] N. H. Bingham, C. M. Goldie, and J. L. Teugels. *Regular Variation*. Number no. 1 in Encyclopedia of Mathematics and its Applications. Cambridge University Press, 1989. MR1015093

[11] St. Boucheron and M. Thomas. Tail index estimation, concentration and adaptivity. *Electronic Journal of Statistics*, 9(2):2751–2792, 2015. MR3435810

[12] V. Brazauskas and R. Serfling. Robust estimation of tail parameters for two-parameter Pareto and exponential models via generalized quantile statistics. *Extremes*, 3(3):231–249, 2001, 2000. MR1856199

[13] M. Brzezinski. Robust estimation of the Pareto tail index: a Monte Carlo analysis. *Empirical Economics*, 51(1):1–30, 2016.

[14] J. Danielsson, L. de Haan, L. Peng, and C. G. de Vries. Using a bootstrap method to choose the sample fraction in tail index estimation. *J. Multivariate Anal.*, 76(2):226–248, 2001. MR1821820

[15] L. de Haan and A. Ferreira. Extreme Value Theory, An Introduction. *Springer Series in Operations Research and Financial Engineering*. Springer, New York, 2006. MR2234156

[16] H. Drees and E. Kaufmann. Selecting the optimal sample fraction in univariate extreme value estimation. *Stochastic Processes and their Applications*, 75(2):149–172, 1998. MR1632189

[17] D. Dupuis and M.-P. Victoria-Feser. A robust prediction error criterion for Pareto modeling of upper tails. *Canadian Journal of Statistics*, 34(4):639–358, 2006. MR2347050

[18] Ch. Dutang, Y. Goegebeur, and A. Guillou. Robust and bias-corrected estimation of the coefficient of tail dependence. *Insurance Math. Econom.*, 57:46–57, 2014. MR3225326

[19] P. Embrechts, C. Klüppelberg, and T. Mikosch. *Modelling Extremal Events*. Springer-Verlag, New York, 1997. MR1458613

[20] M. Finkelstein, H. G. Tucker, and J. A. Veeh. Pareto tail index estimation revisited. *North American Actuarial Journal*, 10(1):1–10, 2006. MR2328626
[21] Y. Goegebeur, A. Guillou, and A. Verster. Robust and asymptotically unbiased estimation of extreme quantiles for heavy tailed distributions. *Statist. Probab. Lett.*, 87:108–114, 2014. MR3168943
[22] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, W. A. Stahel. Robust statistics: the approach based on influence functions. *Wiley series in probability and mathematical statistics. Probability and mathematical statistics*, 2005. MR0829458
[23] P. Hall. On some simple estimates of an exponent of regular variation. *J. Roy. Stat. Assoc.*, 44:37–42, 1982. Series B. MR0655370
[24] P. Hall and A. H. Welsh. Best Attainable Rates of Convergence for Estimates of Parameters of Regular Variation. *The Annals of Statistics*, 12(3):1079–1084, 1984. MR0751294
[25] P. Hall and A. H. Welsh. Adaptive estimates of parameters of regular variation. *Ann. Statist.*, 13, *The Annals of Statistics*, 12(3):331–341, 1985. MR0773171
[26] B. M. Hill. A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3:1163–1174, 1975. MR0378204
[27] P. J. Huber. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*; 35:73–101, 1964. MR0161415
[28] K. Knight. A simple modification of the Hill estimator with applications to robustness and bias reduction. *Technical Report*. http://www.utstat.utoronto.ca/keith/papers/robusthill.pdf.
[29] The Trimmed Hill Estimator: Robust and adaptive tail inference. *Shiny App*. https://shrijita-apps.shinyapps.io/adaptive-trimmed-hill/.
[30] trHill: Hill estimator for upper truncated data. https://rdrr.io/cran/ReIns/man/trHill.html.
[31] E. L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer. MR1639875
[32] L. Peng and A. H. Welsh. Robust estimation of the generalized Pareto distribution. *Extremes*, 4(1):53–65, 2001. MR1876179
[33] J. Pickands. Statistical inference using extreme order statistics. *Ann. Statist.*, 3:119–131, 1975. MR0423667
[34] S. I. Resnick. Heavy-tail phenomena: Probabilistic and statistical modeling. *Springer Series in Operations Research and Financial Engineering*. Springer, New York, 2007. MR2271424
[35] G. Yuri, V. Planchon, J. Beirlant and O. Robert. Quality Assessment of Pedochemical Data Using Extreme Value Methodology. *Journal of Applied Sciences*, 5, 2005.
[36] B. Vandewalle, J. Beirlant, A. Christmann, and M. Hubert. A robust estimator for the tail index of Pareto-type distributions. *Comput. Stat. Data Anal.*, 51(12):6252–6268, August 2007. MR2408592
[37] B. Vandewalle, J. Beirlant, A. Christmann, and M. Hubert. A Robust Estimator of the Tail Index Based on an Exponential Regression Model. *The-*

ory and Applications of Recent Robust Methods, 367–376, January 2004. MR2088312

[38] M.-P. Victoria-Feser and E. Ronchetti. Robust methods for personal-income distribution models. *Canadian Journal of Statistics*, 22(2):247–258, 1994. MR1295691

[39] J. Zou, R. Davis, and G. Samorodnitsky. Extreme Value Analysis Without the Largest Values: What Can Be Done? *Technical Report*. https://people.orie.cornell.edu/gennady/techreports/StrangeHill.pdf.