Journal of Nonparametric Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/gnst20

Asymptotic and finite sample properties of Hill-type estimators in the presence of errors in observations

Mihyun Kim & Piotr Kokoszka

To cite this article: Mihyun Kim & Piotr Kokoszka (2022): Asymptotic and finite sample properties of Hill-type estimators in the presence of errors in observations, Journal of Nonparametric Statistics, DOI: 10.1080/10485252.2022.2136662

To link to this article: https://doi.org/10.1080/10485252.2022.2136662

JOURNAL OF NONPARAMETRIC STATISTICS
https://doi.org/10.1080/10485252.2022.2136662

Asymptotic and finite sample properties of Hill-type estimators in the presence of errors in observations

Mihyun Kim[a] and Piotr Kokoszka[b]

[a]Department of Mathematics, Statistics and Data Science, West Virginia University, Morgantown, WV, USA; [b]Department of Statistics, Colorado State University, Fort Collins, CO, USA

ABSTRACT

We establish asymptotic and finite sample properties of the Hill and Harmonic Moment estimators applied to heavy-tailed data contaminated by errors. We formulate conditions on the errors and the number of upper order statistics under which these estimators continue to be asymptotically normal. We specify analogous conditions which must hold in finite samples for the confidence intervals derived from the asymptotic normal distribution to be reliable. In the large sample analysis, we specify conditions related to second-order regular variation and divergence rates for the number of upper order statistics, k, used to compute the estimators. In the finite sample analysis, we examine several data-driven methods of selecting k, and determine which of them are most suitable for confidence interval inference. The results of these investigations are applied to interarrival times of internet traffic anomalies, which are available only with a round-off error.

ARTICLE HISTORY
Received 11 January 2022
Accepted 6 October 2022

KEYWORDS
Asymptotic normality; harmonic moment estimator; Hill estimator; measurement error; regular variation

2010 MATHEMATICS SUBJECT CLASSIFICATIONS
62F12; 62P30

1. Introduction

Heavy-tailed phenomena have been found in a variety of fields, including finance, insurance, computer network traffic and geophysics. The theory of regular variation provides a mathematical framework for their analysis. Hundreds of papers have been written on the subject, and it is difficult to present an unbiased selection of the most important contributions, so we merely cite here the book of Resnick (2007), and discuss the most closely related references, as the presentation progresses.

This work is concerned with semiparametric estimation of the tail index, α, of a heavy-tailed distribution from observations contaminated by measurement or other errors. We investigate asymptotic and finite sample properties of the Hill estimator, which is the most commonly used tool for inference on α, and of the harmonic moment estimator (HME), which is a class of estimators related to and generalising the Hill estimator. The asymptotic theory establishes conditions on the errors and the number of the largest order statistics, k, that guarantee consistency and asymptotic normality. Finite sample investigation finds the

CONTACT Piotr Kokoszka piotr.kokoszka@colostate.edu Department of Statistics, Colorado State University, Fort Collins, CO 80523, USA

Supplemental data for this article can be accessed here. https://doi.org/10.1080/10485252.2022.2136662

best methods of constructing confidence intervals for α, focusing on data-driven methods for the selection of k, in scenarios where data are observed with errors. While the estimators considered in the paper, especially the Hill estimator, have been extensively explored, their properties in the presence of errors have been mostly unknown.

Suppose $\{X_i, i \geq 1\}$ is a sequence of independent, nonnegative random variables with common distribution function F, which has regularly varying tail probabilities, i.e.

$$\bar{F}(x) = 1 - F(x) = P(X_i > x) = x^{-\alpha} L(x), \quad \alpha > 0, \tag{1}$$

where L is a slowly varying function. The class of distributions with tail behaviour (1) coincides with the maximum domain of attraction of the Fréchet distribution, one of the three basic types of extreme value distributions. The Hill estimator is defined as

$$H_{k,n} = \frac{1}{k} \sum_{i=1}^{k-1} \log \frac{X_{(i)}}{X_{(k)}},$$

with the convention that $X_{(i)}$ is the ith largest order statistic. Throughout the paper, we assume that

$$n \to \infty, \; k \to \infty, \quad \frac{k}{n} \to 0. \tag{2}$$

The Hill estimator is often used after an examination of the Hill plot, which is also a tool for detecting the presence of heavy tails. The Hill plot and the Hill estimator have been extensively studied, and are introduced in all monographs on extreme value theory, see e.g. Embrechts, Klüppelberg, and Mikosch (1997), Beirlant, Goegebeur, Segers, and Teugels (2004), de Haan and Ferreira (2006), Resnick (2007) and Markovich (2008). Considerable research has been done to establish conditions for the asymptotic normality of the Hill estimator. If only the regular variation (1) is assumed, asymptotic normality holds with random centring. Several authors formulated conditions on F, which permit replacing the random centring by a deterministic one. The first result of this type was established by Hall (1982) for slowly varying functions, L, which converge to a constant at a polynomial rate. Davis and Resnick (1984) showed that the estimator is asymptotically normal for any regularly varying function satisfying the von Mises condition, their centring, however, depends on the sample size n. To show that the Hill estimator centred by the exponent α^{-1} is asymptotically normal, second-order regular variation, a refinement of the concept of regular variation, is assumed, see Haeusler and Teugels (1985), Csörgő, Deheuvels, and Mason (1985), Resnick and Stărică (1997a, 1997b). The approach in Section 9.1 of Resnick (2007), which is based on tail empirical processes, also requires the second-order regular variation. Kulik and Soulier (2011) also use the tail empirical process to study asymptotic normality of the Hill estimator for long memory stochastic volatility models assuming a second-order condition.

The HME was introduced by Henry (2009) to provide a broad class of estimators, which, in a sense, extend the Hill estimator and have desirable robustness properties against large outliers. Consistency and asymptotic normality of the HME was established by Henry (2009) for the Pareto distribution and by Beran, Schell, and Stehlík (2014) under a second-order regular variation condition. The HME was also studied, under a different name, by Brilhante, Gomes, and Pestana (2013), Paulauskas and Vaičiulis (2013) and

Caeiro, Gomes, Beirlant, and de Wet (2016). The HME is defined in Beran et al. (2014) by

$$H_{k,n}^{(\beta)} := \frac{1}{\beta - 1} \left\{ \left[\frac{1}{k} \sum_{i=1}^{k} \left(\frac{X_{(k)}}{X_{(i)}} \right)^{\beta - 1} \right]^{-1} - 1 \right\},$$

where $\beta > 0$, $\beta \neq 1$, is a tuning parameter. For $\beta = 1$, the HME is defined by $H_{k,n}^{(1)} := \lim_{\beta \to 1} H_{k,n}^{(\beta)}$. We, therefore, obtain the Hill estimator as the limit of the HME as $\beta \to 1$.

We study the Hill estimator and the HME computed from observations contaminated by measurement errors, or other errors whose origin is either difficult to understand and model or to quantify precisely. We thus assume that we observe

$$Y_i = X_i + \varepsilon_i, \quad 1 \leq i \leq n,$$

where the ε_i are i.i.d. random errors independent of the X_i. The Hill estimator computed from the observations Y_i is then

$$\widehat{H}_{k,n} := \frac{1}{k} \sum_{i=1}^{k-1} \log \frac{Y_{(i)}}{Y_{(k)}},$$

and the HME based on the Y_i is

$$\widehat{H}_{k,n}^{(\beta)} := \frac{1}{\beta - 1} \left\{ \left[\frac{1}{k} \sum_{i=1}^{k} \left(\frac{Y_{(k)}}{Y_{(i)}} \right)^{\beta - 1} \right]^{-1} - 1 \right\}.$$

In our context, $\widehat{H}_{k,n}, \widehat{H}_{k,n}^{(\beta)}$ are the estimators that can be actually used since what we observe are the Y_i, not the X_i. The consistency of the Hill estimator $\widehat{H}_{k,n}$ has been established in very general scenarios in Kim and Kokoszka (2020). In this paper, we want to find conditions under which the asymptotic normality of $\widehat{H}_{k,n}, \widehat{H}_{k,n}^{(\beta)}$ continues to hold. If the errors ε_i have lighter tails than the X_i, the Y_i inherit the regular variation of the X_i. However, the second-order regular variation, needed for the asymptotic normality, is not inherited and suitable conditions that quantify the interplay between the X_i, the ε_i and k must be found. Some specific questions we seek to answer are as follows. What must we assume about the errors ε_i to obtain asymptotic normality with random centring? What additional assumptions are needed for the deterministic centring? In either case, are any additional assumptions on the rate of k, beyond (2), needed? Which characteristics of the distribution of the ε_i enter into these assumptions? In finite samples, how 'large', and in what sense, can the ε_i be for the asymptotic confidence intervals to remain useful? It is hoped that the research we present answers such questions in a useful and informative way.

The problem of estimation in the presence of errors has received considerable attention. For example, Hall and Simar (2002), Goldenshluger and Tsybakov (2004), Kneip, Simar, and Keilegom (2015), and Leng, Peng, Zhou, and Wang (2018) study estimation of the end-point of data observed with additive measurement errors. While they all show asymptotic normality in the presence of Gaussian measurement errors, in our case we assume a broader class of error distributions because the heavy-tailed X_i are 'much larger' random variables than those with a finite end-point. Most closely related is the work of Matsui,

Mikosch, and Tafakori (2013) who study the Hill estimator assuming that the observations have the form $Y_i = 10^{-l}[10^l U_i^{-1/\alpha}]$, where U_i is uniform on $[0,1]$ and $[\cdot]$ denotes the integer part, for $l = 0, 1, 2, \ldots$. Such data can be written in the form of $Y_i = X_i + \varepsilon_i$, where $X_i = U_i^{-1/\alpha}$ has the *exact* Pareto distribution and $\varepsilon_i = 10^{-l}[10^l U_i^{-1/\alpha}] - U_i^{-1/\alpha} \in [-10^{-l}, 0]$ is a non-positive, bounded error of a specific form. Another related, very recent work is Ma, Yan, and Zhang (2022) where rounded data from generalised Pareto distributions are treated as interval-censored data. The parameters of the GPD are estimated by maximum likelihood methods. This method works well in a parametric setting.

We consider broader classes for both the X_i and the ε_i under the assumption that ε_i is independent of X_i, reflecting our treatment of the ε_i as measurement errors. We use a different asymptotic approach. We establish weak convergence of suitable empirical tail processes for observations contaminated by general errors. Asymptotic normality follows from these general results, which are also of independent interest.

The paper is organised as follows. Assumptions and main theoretical results are stated in Section 2. In Section 3, we present simulation studies examining finite sample properties of confidence intervals based on the asymptotic normal distribution, focusing on the impact of errors. This numerical investigation is followed in Section 4 by an application to the interarrival times of internet traffic anomalies. The proofs are presented in Section B of online Supplementary Material, after some preparation in Section A. Additional Tables examining the finite sample performance of the estimators we study are collected in Section C of the online material.

2. Assumptions and main asymptotic results

Recall that the observations are $Y_i = X_i + \varepsilon_i$, $1 \le i \le n$. We first state assumptions on the unobservable random variables X_i. Recall that a function $U : \mathbb{R}_+ \to \mathbb{R}_+$ is regularly varying with index $-\alpha$, $\alpha > 0$, denoted $U \in RV_{-\alpha}$, if

$$\lim_{t\to\infty} \frac{U(tx)}{U(t)} = x^{-\alpha}, \quad \text{for any } x > 0.$$

Assumption 2.1 (Regular variation): The X_i are nonnegative, independent random variables with common distribution function F_X such that $\bar{F}_X = P(X_i > \cdot) \in RV_{-\alpha}$.

Regular variation is not enough to establish asymptotic normality with centring by $1/\alpha$. For this, second-order regular variation is typically assumed. We stated Assumption 2.1 because it is sufficient for certain weaker results that are needed to establish our main result.

Assumption 2.2 (Second-order regular variation (2RV)): The X_i are nonnegative, independent random variables with common distribution function F_X, which is second-order $(-\alpha, \rho)$ regularly varying (written $\bar{F}_X \in 2RV(-\alpha, \rho)$), i.e. there exists a positive function $g \in RV_\rho$ such that $g(t) \to 0$, as $t \to \infty$, and for $\alpha > 0$, $\rho \le 0$, $K \neq 0$,

$$\lim_{t\to\infty} \frac{1}{g(t)} \left(\frac{\bar{F}_X(tx)}{\bar{F}_X(t)} - x^{-\alpha} \right) = H(x) := Kx^{-\alpha} \frac{x^\rho - 1}{\rho}, \quad x > 0. \tag{3}$$

Note that Assumption 2.2 implies Assumption 2.1. Observe, however, that condition (3) does not hold if the X_i have the exact Pareto distribution, i.e. $P(X_i > x) = x^{-\alpha}$. In this

case, one would need to allow $K = 0$, and would thus lose any information contained in the function g. The case of exact Pareto tails should however be included in any reasonable theory for heavy-tailed observations. We do so by introducing a parallel set of assumptions.

Assumption 2.3 (Pareto): The X_i are nonnegative, independent random variables with a common distribution function F_X such that $\bar{F}_X(x) = P(X_i > x) = x^{-\alpha}$, $x \geq 1$, $\alpha > 0$.

The function g in (3) can be interpreted as the convergence rate of $\bar{F}_X(tx)/\bar{F}_X(t)$ to $x^{-\alpha}$. It has been used to restrict the sequence $k = k(n)$. Haeusler and Teugels (1985), Csörgő et al. (1985), Resnick and Stărică (1997a, 1997b) assume that

$$\sqrt{k}g(b(n/k)) \to 0, \tag{4}$$

along with the second-order regular variation for $\rho \leq 0$. In (4), and throughout the paper, $b(\cdot)$ is the quantile function, defined by

$$P(X_i > b(t)) = t^{-1}.$$

It has the representation

$$b(t) = t^{1/\alpha}L_b(t), \tag{5}$$

where L_b is a slowly varying function. Condition (4) is sufficient in our setting if $\rho > -1$. To cover the 2RV case with $\rho \leq -1$ and the pure Pareto case, we consider the following condition:

$$\frac{\sqrt{k}}{b(n/k)} \to 0. \tag{6}$$

Using (5), it is easy to verify that (4) implies $k = o(n^{-2\rho/(\alpha-2\rho)})$, and (6) implies $k = o(n^{2/(\alpha+2)})$. These two rates agree at the phase transition point $\rho = -1$. We use Assumption 2.4 in the 2RV case and Assumption 2.5 in the Pareto case.

Assumption 2.4 (2RV): The sequence $k = k(n)$ satisfies (4) if $\rho > -1$ and (6) if $\rho \leq -1$.

Assumption 2.5 (Pareto): The sequence $k = k(n)$ satisfies (6).

We now turn to the assumptions on the measurement errors ε_i.

Assumption 2.6: The ε_i are i.i.d. with tails satisfying

$$P(|\varepsilon| > x) = o(P(X > x)), \quad \text{as } x \to \infty.$$

The sequence $\{\varepsilon_i\}$ is independent of the sequence $\{X_i\}$.

Under Assumption 2.1, Assumption 2.6 implies that $Y_i = X_i + \varepsilon_i \in RV_{-\alpha}$. It however does not imply that the Y_i satisfy analogs of Assumptions 2.2 or 2.3. To obtain the asymptotic normality with a constant centring, a stronger, but still broadly applicable assumption on the errors is needed; the errors must have lighter tails than a power function. Assumption 2.7 is needed when we assume the second-order regular variation, and Assumption 2.8 is suitable for the Pareto distribution.

Assumption 2.7 (2RV): The ε_i satisfy Assumption 2.6 and

$$P(|\varepsilon| > x) = o(x^{-\kappa}), \quad \text{as } x \to \infty, \tag{7}$$

for some $\kappa > \alpha + \max(-\rho, 1)$.

Assumption 2.8 (Pareto): The ε_i satisfy Assumption 2.6 and (7) for some $\kappa > \alpha + 1$.

We now proceed to define the function spaces in which our functional convergence results hold. We work in $D[0,\infty)$, the Skorokhod space of real-valued, right-continuous functions on $[0,\infty)$ with finite left limits existing on $(0,\infty)$. For any $s>0$, the Skorokhod metric in $D[0,s]$ is defined by

$$d_s(x,y) = \inf_{\lambda\in\Lambda_s} \|\lambda - e\|_s \vee \|x - y\circ\lambda\|_s, \quad x,y \in D[0,s],$$

where $\Lambda_s = \{\lambda : [0,s] \mapsto [0,s],\ \lambda(0)=0,\ \lambda(s)=s,\ \lambda(\cdot) \text{ is continuous, strictly increasing}\}$, and $\|x-y\|_s = \sup_{0\le t\le s}|x(t)-y(t)|$. The Skorokhod metric on $D[0,\infty)$ is then defined by

$$d_\infty(x,y) = \int_0^\infty e^{-s}(d_s(r_s x, r_s y) \wedge 1)\,\mathrm{d}s, \quad x,y \in D[0,\infty),$$

where $r_s x, r_s y$ are the restrictions of $x, y \in D[0,\infty)$ to the interval $[0,s]$. Given a sequence of random processes, $X_n, n \ge 0$, in $D[0,\infty)$, we denote weak convergence of X_n to X_0 by $X_n \Rightarrow X_0$. We also use $\Rightarrow$ to denote weak convergence of random variables.

We define two 'increasingly empirical' measures, with only the last one being observable. We set

$$\nu_n := \frac{1}{k}\sum_{i=1}^n I_{Y_i/b(n/k)}, \quad \hat{\nu}_n := \frac{1}{k}\sum_{i=1}^n I_{Y_i/Y_{(k)}},$$

with $b(\cdot)$ defined in (5). The random measures ν_n, $\hat{\nu}_n$, and all other Radon measures of this type are defined on $(0,\infty]$ compactified at ∞. Thus, for $s \ge 0$, we can define the random processes

$$W_n(s) = \sqrt{k}(\nu_n(s^{-1/\alpha},\infty] - E\nu_n(s^{-1/\alpha},\infty]),$$
$$\widehat{W}_n(s) = \sqrt{k}(\hat{\nu}_n(s^{-1/\alpha},\infty] - E\hat{\nu}_n(s^{-1/\alpha},\infty]).$$

We first investigate the asymptotic normality of the tail empirical processes W_n, $\widehat{W}_n$, then study when it implies the asymptotic normality of the Hill estimator $\widehat{H}_{k,n}$ and the HME $\widehat{H}^{(\beta)}_{k,n}$. Theorem 2.1 shows that even very general errors specified in Assumption 2.6 do not impact the asymptotic behaviour of the tail empirical processes W_n nor $\widehat{W}_n$: the limit distributions of these statistics based on the Y_i are the same as those of the corresponding statistics based on the unobservable X_i.

Theorem 2.1: *Under Assumptions* 2.1 *and* 2.6,

$$W_n \Rightarrow W \quad \text{in } D[0,\infty), \tag{8}$$

and

$$\widehat{W}_n \Rightarrow W \quad \text{in } D[0,\infty), \tag{9}$$

where W is the standard Brownian motion on $[0,\infty)$.

The Hill estimator can be written as an integral of the tail empirical measure $\hat{\nu}_n$, i.e.

$$\widehat{H}_{k,n} = \int_1^\infty \frac{1}{k} \sum_{i=1}^{n} I_{Y_i/Y_{(k)}}(s, \infty] s^{-1} \, \mathrm{d}s = \int_1^\infty \hat{\nu}_n(s, \infty] s^{-1} \, \mathrm{d}s.$$

Similarly, the HME can be expressed as a transformed integral of the tail empirical measure $\hat{\nu}_n$, i.e.

$$\widehat{H}_{k,n}^{(\beta)} = \frac{1}{\beta - 1} \{[(1-\beta)\widehat{M}_{k,n} + 1]^{-1} - 1\}, \quad \beta \neq 1,$$

where

$$\widehat{M}_{k,n}^{(\beta)} := \int_1^\infty \hat{\nu}_n(s, \infty] s^{-\beta} \, \mathrm{d}s = \frac{1}{1-\beta} \left[\frac{1}{k} \sum_{i=1}^{k} \left(\frac{Y_{(k)}}{Y_{(i)}} \right)^{\beta - 1} - 1 \right].$$

The order statistics used to compute the Hill estimator and the HME must be positive. In the following, all statements are tacitly assumed to hold conditional on the event $\{Y_{(k)} > 0\}$, where k is the count of the largest order statistics in the definition of $\widehat{H}_{k,n}$, $\widehat{H}_{k,n}^{(\beta)}$.

Theorem 2.2: *Suppose that Assumptions* 2.1 *and* 2.6 *hold. If* $\alpha > 0$ *and* $\beta > 1 - \alpha/2$,

$$\sqrt{k} \left(\int_1^\infty \hat{\nu}_n(s, \infty] s^{-\beta} \, \mathrm{d}s - \int_{Y_{(k)}}^\infty \frac{n}{k} \bar{F}_Y(s) s^{-\beta} \, \mathrm{d}s \right) \Rightarrow \frac{1}{\alpha} \int_0^1 W(s) s^{\frac{\beta - 1}{\alpha} - 1} \, \mathrm{d}s.$$

By putting $\beta = 1$ in Theorem 2.2 we obtain the asymptotic normality of the Hill estimator with random centring, which is stated as Corollary 2.1(a). Similarly, the asymptotic behaviour of $\widehat{M}_{k,n}^{(\beta)}$ follows directly from Theorem 2.2, which is presented in Corollary 2.1(b).

Corollary 2.1: *Under the assumptions of Theorem* 2.2,

(a)

$$\sqrt{k} \left(\widehat{H}_{k,n} - \int_{Y_{(k)}}^\infty \frac{n}{k} \bar{F}_Y(s) \frac{\mathrm{d}s}{s} \right) \Rightarrow \frac{1}{\alpha} \int_0^1 W(s) \frac{\mathrm{d}s}{s},$$

(b) if $\beta \neq 1$, *then*

$$\sqrt{k} \left(\widehat{M}_{k,n}^{(\beta)} - \int_{Y_{(k)}}^\infty \frac{n}{k} \bar{F}_Y(s) \frac{\mathrm{d}s}{s^\beta} \right) \Rightarrow \frac{1}{\alpha} \int_0^1 W(s) s^{\frac{\beta - 1}{\alpha} - 1} \, \mathrm{d}s.$$

We emphasise that Theorems 2.1, 2.2, and Corollary 2.1 hold either under Assumption 2.2 or Assumption 2.3, since both imply Assumption 2.1.

The convergence in Theorem 2.2 requires random centring with $\int_{Y_{(k)}}^\infty n/k \bar{F}_Y(s) s^{-\beta} \, \mathrm{d}s$, which makes Corollary 2.1 of limited practical use, but it provides a starting point for improvements. To replace it with a constant centring, we need the assumption of second-order regular variation (or of exact Pareto tails) and the stronger assumptions on the errors.

In the following theorem, we establish the asymptotic normality of the integral of the tail empirical measure, $\int_1^\infty \hat{\nu}_n(s,\infty]s^{-\beta}\,ds$, with a constant centring.

Theorem 2.3: *Suppose that Assumptions* 2.2, 2.4, *and* 2.7 *(2RV case) hold. If* $\rho = -1$, *assume, in addition that the limit* $\lim_{t\to\infty} tg(t)$ *exists. Then for any* $\alpha > 1$, $\beta > 1 - \alpha/2$,

$$\sqrt{k}\left(\int_1^\infty \hat{\nu}_n(s,\infty]s^{-\beta}\,\mathrm{d}s - \frac{1}{\alpha+\beta-1}\right) \Rightarrow N\left(0, \frac{\alpha}{(\alpha+\beta-1)^2(\alpha+2\beta-2)}\right). \quad (10)$$

Under Assumptions 2.3, 2.5, *and* 2.8 *(Pareto case),* (10) *holds for* $\alpha > 0$, $\beta > 1 - \alpha/2$.

Remark 2.1: The case of the second-order regular variation exponent $\rho = -1$ needs special treatment because our arguments require that $\lim_{t\to\infty} tg(t)$ exists (∞ is allowed). By Proposition 2.6(i) in Resnick (2007), if $\rho > -1$, then $\lim_{t\to\infty} tg(t) = \infty$, and if $\rho < -1$, then $\lim_{t\to\infty} tg(t) = 0$. If $\rho = -1$, $\lim_{t\to\infty} tg(t)$ need not exist.

The asymptotic normality of the Hill estimator $\widehat{H}_{k,n}$ follows easily from Theorem 2.3. To obtain the asymptotic normality of the HME $\widehat{H}^{(\beta)}_{k,n}$, we must apply Theorem 2.3 and the delta method. The corresponding results are stated in the following corollary.

Corollary 2.2: *Under the assumptions of Theorem* 2.3,

(a)

$$\sqrt{k}\left(\widehat{H}_{k,n} - \frac{1}{\alpha}\right) \Rightarrow N(0, 1/\alpha^2), \quad (11)$$

(b) if $\beta \neq 1$, *then*

$$\sqrt{k}\left([(1-\beta)\widehat{M}^{(\beta)}_{k,n} + 1] - \frac{\alpha}{\alpha+\beta-1}\right) \Rightarrow N\left(0, \frac{\alpha(1-\beta)^2}{(\alpha+\beta-1)^2(\alpha+2\beta-2)}\right)$$

and

$$\sqrt{k}\left(\widehat{H}^{(\beta)}_{k,n} - \frac{1}{\alpha}\right) \Rightarrow N\left(0, \frac{(\alpha+\beta-1)^2}{\alpha^3(\alpha+2\beta-2)}\right). \quad (12)$$

The limits in (11) and (12) are the same as for observations without measurement errors; see Theorem 3.2.5 of de Haan and Ferreira (2006) and Theorem 2 of Beran et al. (2014). The effect of suitably small errors ε_i is thus asymptotically negligible. However, even for such errors, we impose conditions (4) and (6) on the rate of k in the cases of 2RV ($\rho \leq -1$) and exact Pareto observations, respectively. We do not know if Corollary 2.2 remains true without these conditions on k. We also remark that Corollary 2.2(a) cannot be easily proven by verifying the conditions in Theorem 3.2.5 of de Haan and Ferreira (2006). If the X_i are exactly Pareto or second-order regularly varying, the Y_i need not be in any of these classes. Proposition B.1 in the online material, which may be useful in other contexts, is a related result which plays an important role in the proof of Theorem 2.3.

In the next two sections, we explore how small the errors must be in finite samples to have a practically negligible effect on confidence interval inference.

3. Impact of errors on confidence intervals

We investigate the effect of error contaminations on confidence intervals constructed using the more commonly used Hill estimator. The effect of various errors on the harmonic moment estimator (HME) is studied in a more limited, but informative, simulation study presented in Section C.2 of the online material.

The asymptotic level $1-p$ confidence interval for α^{-1} implied by Corollary 2.2 (a) is

$$\left(\frac{1}{\hat{\alpha}} - z_{p/2}\frac{1}{\hat{\alpha}\sqrt{k}},\ \frac{1}{\hat{\alpha}} + z_{p/2}\frac{1}{\hat{\alpha}\sqrt{k}}\right), \tag{13}$$

where $\hat{\alpha}^{-1} = \widehat{H}_{k,n}$, and z_q is the upper quantile of the standard normal distribution defined by $\Phi(z_q) = 1 - q$. The above interval is implemented by the function `hill` of the `R` package `evir`, with the default asymptotic coverage $1-p = 0.95$. According to Corollary 2.2(a), it is asymptotically valid even if the observations are contaminated by fairly general errors. In this section, we investigate the impact of these errors on the empirical coverage probability of the interval (13). To obtain interval (13), the number of upper order statistics, k, has to be chosen. We consider a range of values of k for a given sample size n. We also employ a few methods of selecting k, which have been proposed.

The design of our simulation study is as follows. We generate observations $Y_i = X_i + \varepsilon_i$, $i = 1, 2, \ldots, n$, where $\{X_i\}$ and $\{\varepsilon_i\}$ are independent sets of random variables. For each model/error pair, we compute 1000 confidence intervals and report the fraction of the intervals that contain the reciprocal of the true tail index. We consider sample sizes $n = 500$ and $n = 2000$. The sample size $n = 500$ is representative of the sample sizes occurring in the application presented in Section 4.

We use two models for the X_i, both satisfying the condition of Corollary 2.2(a) and having the true tail index $\alpha = 2$. The first is the standard Pareto distribution, which is not second-order regularly varying, and the second is a distribution in the Hall/Weiss class. The Hall/Weiss class provides examples of the second-order regular variation, see p. 142 of Geluk, de Haan, Resnick, and Stărică (1997). Model 2 satisfies Assumption 2.2 with $g(t) = t^{-5}$.

Model 1 [Pareto] The X_i are i.i.d. random variables, which follow a Pareto distribution with $\alpha = 2$, $P(X_i > x) = x^{-2}$, $x \geq 1$.

Model 2 [2RV] The X_i are i.i.d. random variables, which follow the Hall/Weiss class with $\alpha = 2$ and $\rho = -5$, $P(X_i > x) = x^{-2}(1 + x^{-5})/2$, $x \geq 1$.

We consider four different distributions for the errors ε_i. They all satisfy Assumptions 2.7 and 2.8 (with $\alpha = 2$), since for each of them $P(|\varepsilon| > x) = o(x^{-\kappa})$, for some $7 < \kappa < 8$.

Error 1 [Normal] The ε_i are i.i.d. random variables, drawn from a normal distribution with mean 0 and standard deviation σ_{Normal}.

Error 2 [scaled t_8] The ε_i are i.i.d. random variables, drawn from a scaled t-distribution with 8 degrees of freedom.

Error 3 [GPD] The ε_i are i.i.d. random variables, drawn from a generalised Pareto distribution, $P(|\varepsilon| > z) = (1 + \xi(z - \mu)/\sigma)^{-1/\xi}$, with location $\mu = 0$, shape $\xi = 1/8$, and scale σ_{GPD}.

Error 4 [Uniform] The ε_i are i.i.d. random variables, drawn from the uniform distribution on the interval $[-a, a]$, $a > 0$.

In the investigations that follow, we need to separate the effect of the shape of the density from the effect of the typical size of the error relative to the size of the X_i. We do so by reporting the ratio of the *sample* SDs: (error SD)/(model SD). The X_i we consider have infinite variance, but the sample SD is always finite and provides a measure of the size of the generated data.

We first consider a wide range of k for a given sample size n. Tables 5 and 6 in Section C of the online Supplementary Material report coverage probabilities of the approximate 95% confidence intervals for the Pareto model, with $n = 500$ and $n = 2000$, respectively. We first observe that the coverage probabilities for samples generated from the Pareto distribution without the errors are close to the target coverage, 95%, for large k's. This is found in the row with the ratio 0 in each table. This result is in agreement with the typical behaviour of the Hill plot showing stable, unbiased estimates for large k when its underlying distribution is exactly a Pareto distribution. Second, the coverage overall decreases with the ratio, but this decrease is relatively flat over a range of the ratio from 0.01 to 0.1, for all the error types. In particular, for $n = 2000$, the coverage is surprisingly acceptable for a wide range of values of k; in many cases, it is close to the target of 95%. On the other hand, the coverage seems sensitive to relatively large errors with a ratio of more than 10%. An interesting observation is that, in the presence of errors, the coverage gets worse as k gets larger. This result is consistent with Corollary 2.2(a), which implies that the Hill estimator obtains the asymptotic normality if k satisfies Assumption 2.5; k goes to infinity with n, but not too fast. The reduction in the coverage probability caused by large k is not observed for data contaminated by relatively small errors. Finally, the impact on the coverage probability overall does not depend on the type of error distribution. In particular, for the small ratios, the difference that the error type makes looks negligible.

Tables in Section C of Supplementary Material report coverage probabilities of the asymptotic 95% confidence intervals for the 2RV model, with $n = 500$ and $n = 2000$, respectively. Unlike the Pareto case, the 2RV model does not achieve the target coverage, 95%, even if there are no errors. This may be due to n not being sufficiently large. The errors with a small ratio, however, have only a small impact on the coverage. It can be also seen that the impact on the coverage probability for a small ratio does not depend on the error type. Finally, we see that k cannot increase too fast, indirectly confirming the need for Assumption 2.4.

We have found so far that the coverage can achieve the target probability for some properly chosen k or cannot achieve it for any k, given a finite sample. Even if we can identify some range of k for which the coverage approaches the target, the question still remains of how to select an optimal k in practice. There are various methods for choosing it. A commonly used approach is based on the minimisation of the asymptotic mean squared error (AMSE), see e.g. Hall and Welsh (1985), Hall (1990), Drees and Kaufmann (1998), and Danielsson, de Haan, Peng, and de Vries (2001). These methods are however based on asymptotic arguments, which brings up a question of how well they perform in finite samples. Danielsson, Ergun, de Haan, and de Vries (2019) suggest a data-driven method minimising a penalty function of the distance between empirical quantiles and theoretical quantiles to improve the performance in finite samples. There are also heuristic methods, mainly trying to find the region where the Hill plot, a plot of estimates of the tail index against k, becomes more stable, see Resnick and Stărică (1997b).

To provide practically useful information on choosing a data-driven cut-off k, we examined four methods based on different underlying ideas of selecting the optimal k. The first threshold selection method, introduced by Hall (1990), uses a bootstrap procedure to find the k which minimises the AMSE. This value is computed by the function `hall` of the `R` package `tea`. (We also considered a few related methods based on the minimisation of the AMSE argument, but they all gave disappointing results. The coverage that the Hall method produced was always among the best of these methods.) The second method, proposed by Danielsson et al. (2019), is based on minimising a penalty function of the distance between the observed quantile and the fitted Pareto-type tail. This distance is in the quantile dimension, not in the probability dimension like the Kolmogorov–Smirnov distance. This method is suggested to remedy the behaviour that a small change in probabilities makes a large difference in quantiles. We use two different penalty functions: the supremum of the absolute distance (KS), and the mean absolute distance (MAD). Both are implemented by the function `mindist` of the `R` package `tea`. The final method is an Eye–Ball technique whose automatic algorithm is developed by Danielsson et al. (2019) and is carried out by the function `eye` of the `R` package `tea`. This heuristic method attempts to find a stable portion of the Hill plot and obtain the k at which a considerable drop in the variance occurs, as k increases.

Tables 1 and 2 report coverage probabilities and the average optimal k selected using the four different methods. For the Pareto model, the coverage decreases with the ratio for all the selection methods as shown in Table 1; again, a small ratio has a relatively small impact on the coverage. The MAD and Eye–Ball methods achieve the target coverage, 95%, when the underlying process is not contaminated by the errors. These methods also are less sensitive to the ratio increase. For the Pareto model, the MAD approach generally leads to coverage probabilities which are higher than 95%. However, as shown in Table 2, it gives very low coverage for the 2RV model. It has an unexpected, difficult to explain, property of the coverage increasing with the ratio. The Hall method also shows some fluctuation over the ratio, but this fluctuation is not found when the ratio is 0.01 and 0.02. The other methods also exhibit this insensitivity for small ratios. The Eye–Ball method seems to work well for the Pareto and 2RV models since it gives relatively high values of coverage. Its average optimal k also falls into the optimal range which gives high values of coverage in Tables 5 and 7 in Section C.

The main conclusions of the above-detailed discussion are as follows.

(1) The Eye–Ball method of selecting k is recommended for both the Pareto and 2RV models.
(2) For the heavy-tailed X_i with the tail index $\alpha = 2$, the coverage probability of the approximate 95% confidence interval containing the true index is robust to errors whose SD does not exceed 2% of model SD.
(3) There is no clear evidence that the coverage probability depends on the error distribution. Instead, the coverage is mainly affected by how large the ε_i are compared to the X_i, regardless of the threshold selection methods.

We conclude this section with a discussion of the confidence interval for α obtained via an application of the delta method. Corollary 2.2(a) and the delta method imply

Table 1. Proportion (in percent) of the approximate 95% confidence intervals including $1/\alpha$ and the average optimal k in parentheses, for $n = 500$ and the **Pareto** model.

		Error SD/model SD ratio						
Method	Error type	0	0.01	0.02	0.05	0.1	0.2	0.3
Hall	Normal	88.9	87.6	88.4	88.9	83.8	77.5	71.2
		(283)	(311)	(330)	(329)	(301)	(256)	(222)
	scaled t_8	88.7	88.0	88.6	88.9	83.2	77.6	68.9
		(283)	(321)	(337)	(322)	(289)	(242)	(201)
	GPD	89.4	88.9	88.8	88.7	83.7	76.9	72.7
		(285)	(322)	(340)	(320)	(283)	(220)	(169)
	Uniform	89.1	88.2	88.3	87.9	80.1	73.1	61.3
		(284)	(308)	(329)	(329)	(301)	(265)	(238)
MAD	Normal	97.0	97.4	96.8	97.6	96.8	97.4	96.2
		(218)	(214)	(214)	(198)	(147)	(92)	(68)
	scaled t_8	97.1	97.2	97.4	97.8	97.2	97.2	97.2
		(219)	(219)	(214)	(200)	(156)	(107)	(79)
	GPD	97.1	97.2	98.0	98.2	97.8	98.2	98.0
		(219)	(216)	(214)	(191)	(145)	(95)	(77)
	Uniform	97.0	97.4	96.4	96.2	97.0	94.8	93.6
		(218)	(219)	(220)	(195)	(151)	(99)	(70)
KS	Normal	83.4	82.2	84.0	81.2	77.2	75.0	67.6
		(68)	(67)	(68)	(77)	(93)	(83)	(79)
	scaled t_8	83.6	83.6	83.5	84.2	81.7	77.4	71.9
		(68)	(67)	(69)	(71)	(90)	(85)	(82)
	GPD	83.6	84.4	83.8	83.8	82.4	77.9	75.1
		(68)	(68)	(66)	(72)	(83)	(74)	(62)
	Uniform	83.4	84.0	82.0	82.0	78.6	69.4	63.6
		(68)	(70)	(69)	(80)	(92)	(103)	(101)
Eye	Normal	95.3	95.1	94.8	95.2	94.8	93.2	90.5
		(51)	(51)	(51)	(51)	(51)	(51)	(50)
	scaled t_8	95.3	95.4	95.5	95.3	93.5	92.7	88.2
		(51)	(51)	(51)	(51)	(51)	(50)	(50)
	GPD	95.3	95.2	94.9	95.2	93.5	91.7	86.0
		(51)	(51)	(51)	(51)	(50)	(50)	(49)
	Uniform	95.3	95.1	95.0	95.6	94.3	93.6	92.0
		(51)	(51)	(51)	(51)	(51)	(51)	(51)

Notes: The Hall, MAD, KS, and Eye–Ball methods are used to choose the optimal k. The target coverage is 95%.

that

$$\sqrt{k}(\widehat{H}_{k,n}^{-1} - \alpha) \Longrightarrow N(0, \alpha^2).$$

Thus, setting $\tilde{\alpha} = \widehat{H}_{k,n}^{-1}$, we get the approximate level $1-p$ confidence interval for α of the form

$$\left(\tilde{\alpha} - z_{p/2}\frac{\tilde{\alpha}}{\sqrt{k}},\ \tilde{\alpha} + z_{p/2}\frac{\tilde{\alpha}}{\sqrt{k}}\right). \tag{14}$$

One might want to use the interval (14) rather than (13) to make inference on α, but care is needed in finite samples. Since the delta method is based on an additional asymptotic approximation, confidence intervals derived from it could provide a poor approximation for small sample sizes. We have performed a simulation study for the interval (14), similar to the one described earlier in this section. We have found that it almost always gives coverage probability worse than the interval (13). Therefore, when working with sample sizes similar to $n = 500$ or $n = 2000$, we recommend using the reciprocals of the bounds of the interval (13).

Table 2. Proportion (in percent) of the approximate 95% confidence intervals including $1/\alpha$ and the average optimal k in parentheses, for $n = 500$ and the **2RV** model.

		Error SD/model SD ratio						
Method	Error type	0	0.01	0.02	0.05	0.1	0.2	0.3
Hall	Normal	75.3	75.6	75.0	12.9	8.8	37.2	34.7
		(118)	(119)	(142)	(395)	(416)	(326)	(250)
	scaled t_8	75.8	75.3	74.4	29.0	0.8	29.1	37.4
		(118)	(119)	(130)	(339)	(429)	(366)	(293)
	GPD	75.8	76.2	72.5	28.3	1.9	17.6	38.3
		(118)	(119)	(150)	(340)	(422)	(368)	(284)
	Uniform	75.6	75.2	74.0	33.8	26.4	35.0	30.3
		(118)	(119)	(129)	(316)	(410)	(314)	(256)
MAD	Normal	18.7	18.2	18.5	16.3	7.5	36.6	70.8
		(222)	(221)	(221)	(228)	(304)	(150)	(90)
	scaled t_8	18.7	18.9	18.8	17.0	8.5	21.0	51.7
		(222)	(221)	(222)	(223)	(311)	(200)	(121)
	GPD	18.7	18.3	18.8	16.3	12.1	22.2	66.2
		(222)	(221)	(221)	(223)	(270)	(203)	(127)
	Uniform	18.7	18.4	18.1	16.1	11.1	40.6	60.2
		(222)	(222)	(221)	(240)	(282)	(134)	(82)
KS	Normal	66.6	66.6	67.0	66.4	56.7	52.5	53.0
		(104)	(102)	(104)	(104)	(152)	(160)	(117)
	scaled t_8	66.6	66.9	66.8	67.0	66.3	53.9	53.0
		(104)	(103)	(105)	(100)	(117)	(175)	(136)
	GPD	66.6	66.4	67.5	67.0	65.6	56.8	58.8
		(104)	(103)	(104)	(101)	(109)	(155)	(115)
	Uniform	66.6	67.1	66.8	66.3	53.9	51.0	48.4
		(104)	(102)	(102)	(107)	(169)	(166)	(140)
Eye	Normal	93.6	93.9	93.4	93.7	92.6	88.7	77.8
		(51)	(51)	(51)	(51)	(51)	(51)	(50)
	scaled t_8	93.6	93.9	94.6	93.9	91.9	91.1	83.3
		(51)	(51)	(51)	(51)	(51)	(50)	(50)
	GPD	93.6	93.8	93.8	93.9	92.5	88.1	82.3
		(51)	(51)	(51)	(51)	(51)	(50)	(50)
	Uniform	93.6	93.7	93.8	94.0	92.7	90.4	82.9
		(51)	(51)	(51)	(51)	(51)	(51)	(51)

Notes: The Hall, MAD, KS, and Eye–Ball methods are used to choose the optimal k. The target coverage is 95%.

Finally, we note that a preliminary simulation study indicates that the moment estimator of Dekkers, Einmahl, and de Haan (1989) might also be robust to errors and suitable for the construction of confidence intervals in case of error contaminated data. A separate theoretical and empirical study is needed.

4. Application to Internet2 anomalous traffic

In this section, we present an application to interarrival times of anomalies in a backbone internet network, Internet2. These times are available only with round-off errors. We provide only minimal background; more details are presented in Bandara, Pezeshki, and Jayasumana (2014), a paper which to some extent motivates the present research. We describe the results of confidence interval inference for the tail index of these interarrival times. We restrict ourselves to confidence intervals based on the Hill estimator, the results for the HME are similar. We then examine the robustness of the Hill estimator to the round-off errors by a numerical experiment.

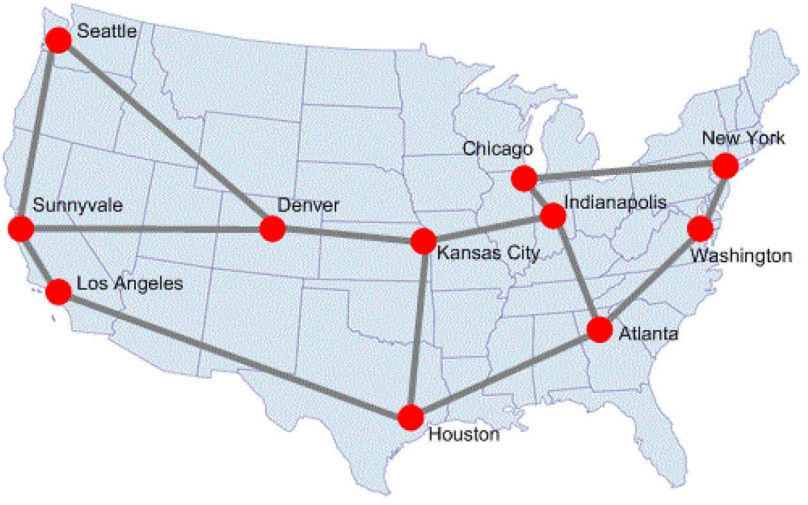

Figure 1. A map showing 14 two-directional links of the Internet2 network.

The Internet2 network consists of 14 two-directional links connecting major cities in the United States, as shown in Figure 1. A traffic disruption in any of these links can slow down service in the whole country. For this reason, anomalies in the internet traffic have been extensively studied. An anomaly is a time and space confined traffic whose volume is much higher than typical. Bandara et al. (2014) developed an anomaly extraction algorithm. The anomaly extraction algorithm can identify the arrival time of an anomaly in any unidirectional link only in a resolution of five minutes. While network measurement devices operate at much higher frequencies, such a rough resolution is due to the limitation of the anomaly extraction algorithm. It is based on the Fourier transform, which eliminates noise by retaining only low-frequency harmonics. Bandara et al. (2014) created a database for the time period of 50 weeks, starting 16 October 2005. A question we seek to answer in this section is if the round-off error has a negligible or a non-negligible impact on the confidence intervals for the tail index of the interarrival times. Additionally, we would like to see if the various data-driven methods of selecting k, discussed in Section 3 lead to overlapping confidence intervals, or if they suggest different ranges of α. These conclusions could potentially be different for each of the 28 unidirectional links. We index these links by integers from 1 to 28 since it is not important for the purpose of our investigation to which nodes they correspond.

In the context of this paper, each interarrival time Y_i, computed by the algorithm, is treated as a 'true' interarrival time X_i measured with a round-off error, i.e. $Y_i = X_i + \varepsilon_i$. The unobserved X_i is not rigorously defined, but we can think of it as the time separation based on a more precise algorithm, or just a different algorithm. In the latter case, the analysis that follows provides information about the uncertainty in the estimation of α caused by the choice of a specific algorithm. The value of the ε_i does not depend on X_i because there is no reason to believe that, say, larger X_i have a 'preference' for falling into some specific part of the 5-minute interval separating the possible measurement times. (The X_i are at least a few hours.) The errors need not be negative and it is risky to assume that the X_i have exact Pareto tails, so the theory of Matsui et al. (2013) does not apply.

Table 3. Point estimates and 95% confidence intervals for the tail index of the anomalies interarrival times.

Link	1 ($n = 405$)		2 ($n = 247$)		3 ($n = 362$)		4 ($n = 454$)	
Hall	1.70	(1.3, 2.4)	1.50	(1.1, 2.5)	1.63	(1.3, 2.1)	1.64	(1.3, 2.1)
MAD	1.43	(1.2, 1.8)	1.21	(1.0, 1.6)	1.51	(1.1, 2.3)	1.28	(0.9, 2.6)
KS	3.19	$(1.3, \infty)$	3.06	(1.6, 24.8)	4.66	$(2.0, \infty)$	2.08	(1.2, 8.0)
Eye	1.79	(1.4, 2.4)	1.23	(1.0, 1.8)	1.60	(1.3, 2.2)	1.59	(1.3, 2.1)
Overlap		(1.4, 1.8)		(1.6, 1.6)		(2.0, 2.1)		(1.3, 2.1)
	5 ($n = 347$)		6 ($n = 345$)		7 ($n = 603$)		8 ($n = 300$)	
Hall	1.54	(1.2, 2.1)	1.59	(1.2, 2.2)	1.64	(1.3, 2.2)	1.45	(1.1, 2.1)
MAD	1.43	(1.2, 1.8)	1.49	(1.0, 2.9)	1.31	(0.9, 2.4)	1.27	(1.0, 1.7)
KS	1.88	(1.3, 3.2)	3.35	(1.9, 12.9)	5.34	(3.2, 17.4)	3.43	(2.1, 9.9)
Eye	1.53	(1.2, 2.1)	1.52	(1.2, 2.1)	1.38	(1.1, 1.7)	1.38	(1.1, 1.9)
Overlap		(1.3, 1.8)		(1.9, 2.1)		$\emptyset$		$\emptyset$
	9 ($n = 387$)		10 ($n = 345$)		11 ($n = 382$)		12 ($n = 304$)	
Hall	1.48	(1.2, 2.1)	1.44	(1.1, 2.1)	1.83	(1.4, 2.6)	2.27	(1.6, 3.7)
MAD	1.31	(1.1, 1.7)	1.24	(1.0, 1.6)	1.36	(1.1, 1.8)	1.50	(1.2, 2.0)
KS	3.98	(2.3, 15.4)	2.85	(1.7, 9.3)	3.63	$(1.7, \infty)$	2.51	(1.7, 4.9)
Eye	1.52	(1.2, 2.0)	1.39	(1.1, 1.9)	1.72	(1.4, 2.3)	1.60	(1.3, 2.2)
Overlap		$\emptyset$		$\emptyset$		(1.7, 1.8)		(1.7, 2.0)
	13 ($n = 476$)		14 ($n = 507$)		15 ($n = 478$)		16 ($n = 319$)	
Hall	2.16	(1.7, 3.0)	1.96	(1.5, 3.0)	2.07	(1.6, 3.0)	1.44	(1.1, 2.0)
MAD	1.58	(1.0, 3.9)	1.44	(0.9, 3.5)	1.46	(0.9, 3.4)	1.36	(1.0, 2.2)
KS	2.06	(1.6, 2.9)	3.85	$(1.3, \infty)$	2.05	(1.6, 2.9)	3.32	(1.9, 16.6)
Eye	2.02	(1.6, 2.7)	1.60	(1.3, 2.1)	1.80	(1.5, 2.3)	1.47	(1.2, 2.0)
Overlap		(1.7, 2.7)		(1.5, 2.1)		(1.6, 2.3)		(1.9, 2.0)

Notes: The link index along with the sample size are displayed. The estimates are obtained using the Hall, MAD, KS, and Eye–Ball methods. The intersection of the four intervals is shown if it is nonempty, an empty intersection is indicated by $\emptyset$.

Kokoszka, Nguyen, Wang, and Yang (2020) and Nicholson, Kokoszka, Lund, Kiessler, and Sharp (2021) showed that the Y_i have regularly varying, but not exact Pareto tails. The autocorrelation analysis in these papers also showed that the Y_i can be assumed to be i.i.d.

Tables 3 and 4 report tail index estimates and 95% confidence intervals for each link, obtained using the four methods of selecting k discussed in Section 3. We first observe that all methods, except for the KS method, generally produce similar point estimates for each link. The interval estimates from the KS method are generally wider. In particular, some links have the infinity as the upper end. This is manually put in to deal with a negative lower end of the interval (13). We now check whether intervals from the four methods overlap. We find 20 links with a nonempty intersection of the 4 intervals and 8 links with an empty intersection. The intersection does not have any interpretation in the usual frequentist sense of Neyman (1937), but it provides, so to say, the safest region in an engineering sense, for the 20 links for which it is nonempty. For the links with the empty intersection, or even for all links, we recommend using the confidence interval produced from the Eye–Ball method, which can be considered the most reliable estimate based on the simulation result of Section 3.

We conclude this section by reporting results of an experiment designed to assess if the rounding-off errors have a practical impact on the estimates of α. For each link, we treat the value of α estimated from the observed interarrival times as the true value and the observed Y_i as the true X_i. We generate $R = 1000$ replications of error contaminated data

Table 4. Continuation of Table 3.

Link	17 ($n = 402$)		18 ($n = 388$)		19 ($n = 433$)		20 ($n = 493$)	
Hall	1.91	(1.5, 2.5)	1.36	(1.1, 1.9)	1.27	(1.1, 1.6)	1.90	(1.5, 2.6)
MAD	1.51	(1.0, 3.7)	1.22	(1.0, 1.6)	1.27	(1.0, 1.9)	1.45	(0.9, 3.3)
KS	1.96	(1.5, 2.8)	3.22	(1.7, 26.1)	2.63	$(1.2, \infty)$	1.97	(1.6, 2.6)
Eye	1.86	(1.5, 2.5)	1.31	(1.1, 1.8)	1.51	(1.2, 2.0)	1.83	(1.5, 2.4)
Overlap		(1.5, 2.5)		$\emptyset$		(1.2, 1.6)		(1.6, 2.4)
	21 ($n = 340$)		22 ($n = 417$)		23 ($n = 597$)		24 ($n = 296$)	
Hall	1.97	(1.5, 2.9)	1.46	(1.2, 1.9)	1.67	(1.3, 2.2)	1.56	(1.2, 2.2)
MAD	1.51	(1.0, 3.3)	1.38	(1.1, 2.0)	1.26	(0.8, 2.9)	1.28	(1.0, 1.7)
KS	2.01	(1.5, 3.0)	3.61	$(1.5, \infty)$	3.67	(2.1, 14.2)	3.44	(2.0, 13.3)
Eye	1.87	(1.5, 2.6)	1.54	(1.2, 2.1)	1.50	(1.2, 1.9)	1.43	(1.1, 2.0)
Overlap		(1.5, 2.6)		(1.5, 1.9)		$\emptyset$		$\emptyset$
	25 ($n = 258$)		26 ($n = 340$)		27 ($n = 348$)		28 ($n = 264$)	
Hall	1.78	(1.3, 2.9)	1.48	(1.1, 2.3)	1.95	(1.5, 2.9)	1.58	(1.2, 2.3)
MAD	1.35	(1.0, 1.9)	1.20	(0.9, 1.7)	1.64	(1.0, 4.7)	1.38	(1.0, 2.4)
KS	4.11	$(1.7, \infty)$	3.57	(2.2, 10.3)	2.80	(1.6, 14.0)	2.70	$(1.3, \infty)$
Eye	1.38	(1.1, 2.0)	1.25	(1.0, 1.7)	1.71	(1.4, 2.4)	1.60	(1.2, 2.3)
Overlap		(1.7, 1.9)		$\emptyset$		(1.6, 2.4)		(1.3, 2.3)

$Y_i^{(r)} = Y_i + \varepsilon_i^{(r)}$, $1 \le r \le 1000$. We assume that the errors are uniformly distributed on $[-1, 1]$, because, as noted above, there is no reason why the X_i should prefer some parts of the 5-minute interval. (The data are normalised so that this interval corresponds to the interval $[0, 1]$.) For each of these replications we compute the interval (13) with $p = 10\%$ and $p = 5\%$. To choose k, we use the Hall, MAD, KS, and Eye–Ball methods described in Section 3. For each link, we determine the percentage of these intervals that cover the value of α estimated from real data. If the interarrival times were measured perfectly, i.e. $\varepsilon_i \equiv 0$, then 100% of these intervals would cover the 'true value', so our target in this experiment is 100% rather than 95% or 90% as in Section 3. If the actual coverage is $100(1 - q)\%$, then we interpret q as the probability of getting a wrong interval estimate due to the round-off error. It turned out that for all links we achieved the target percentage, 100%, for both 95% and 90% confidence levels, regardless of the threshold selection methods. In light of the results of Section 3, the 100% coverage could be expected since the ratio of the Error SD to the observation SD is less than 0.001 for each link. We have seen from Tables 1 and 2 that the errors with the ratio of 0.01 had almost no impact on the coverage probability. Based on this 100% coverage, we conclude that the impact of the round-off error on the confidence interval estimate from the real data is practically negligible. This allows us to use the available rough interarrival times to make an inference on the tail index.

The conclusions of the research described in this section are as follows.

(1) For the purpose of confidence interval inference on the tail index of the anomalies interarrival times, the 5-minute resolution is acceptable.
(2) For most links, the confidence intervals obtained using the four data-driven methods of selecting k have a nonempty intersection.
(3) Based on the Eye–Ball method, one can be confident that for all links the true value of α is between 1.0 and 2.7. The most typical range for α is (1.2, 2.3); each interval for half of the links falls into the range.

Acknowledgments

Section 4 uses a proprietary data product derived from historical US-wide internet traffic measurements. We thank Professor Anura P. Jayasumana of Colorado State University's Department of Electrical and Computer Engineering for making it available to us.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research has been partially supported by the National Science Foundation grant DMS-1737795: *ATD: Spatio-Temporal Model for the Propagation of Internet Traffic Anomalies*.

References

Bandara, V.W., Pezeshki, A., and Jayasumana, A.P. (2014), 'A Spatiotemporal Model for Internet Traffic Anomalies', *IET Networks*, 3, 41–53.

Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J. (2004), *Statistics of Extremes: Theory and Applications*, Chichester: John Wiley & Sons.

Beran, J., Schell, D., and Stehlík, M. (2014), 'The Harmonic Moment Tail Index Estimator: Asymptotic Distribution and Robustness', *Annals of the Institute of Statistical Mathematics*, 66, 193–220.

Brilhante, M., Gomes, M., and Pestana, D. (2013), 'A Simple Generalisation of the Hill Estimator', *Computational Statistics & Data Analysis*, 57, 518–535.

Caeiro, F., Gomes, M., Beirlant, J., and de Wet, T. (2016), 'Mean–of–order P Reduced–bias Extreme Value Index Estimation Under a Third–order Framework', *Extremes*, 19, 561–589.

Csörgő, S., Deheuvels, P., and Mason, D. (1985), 'Kernel Estimates of the Tail Index of a Distribution', *The Annals of Statistics*, 13, 1050–1077.

Danielsson, J., de Haan, L., Peng, L., and de Vries, C. (2001), 'Using a Bootstrap Method to Choose the Sample Fraction in Tail Index Estimation', *Journal of Multivariate Analysis*, 76, 226–248.

Danielsson, J., Ergun, L.M., de Haan, L., and de Vries, C.G (2019), 'Tail Index Estimation: Quantile Driven Threshold Selection', Technical Report, Bank of Canada.

Davis, R.A., and Resnick, S.I. (1984), 'Tail Estimates Motivated by Extreme Value Theory', *The Annals of Statistics*, 12, 1467–1487.

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

Dekkers, A., Einmahl, J., and de Haan, L. (1989), 'A Moment Estimator for the Index of an Extreme-Value Distribution', *The Annals of Statisctis*, 17, 1833–1855.

Drees, H., and Kaufmann, E. (1998), 'Selecting the Optimal Sample Fraction in Univariate Extreme Value Estimation', *Stochastic Processes and their Applications*, 75, 149–172.

Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997), *Modelling Extremal Events for Insurance and Finance*, Berlin: Springer.

Geluk, J., de Haan, L., Resnick, S.I., and Stărică, C. (1997), 'Second-Order Regular Variation, Convolution and the Central Limit Theorem', *Stochastic Processes and their Applications*, 69, 139–159.

Goldenshluger, A., and Tsybakov, A. (2004), 'Estimating the Endpoint of a Distribution in the Presence of Additive Observation Errors', *Statistics & Probability Letters*, 68, 39–49.

Haeusler, E., and Teugels, J.L. (1985), 'On Asymptotic Normality of Hill's Estimator for the Exponent of Regular Variation', *The Annals of Statistics*, 13, 743–756.

Hall, P. (1982), 'On Some Simple Estimates of an Exponent of Regular Variation', *Journal of the Royal Statistical Society: Series B (Methodological)*, 44, 37–42.

Hall, P. (1990), 'Using the Bootstrap to Estimate Mean Squared Error and Select Smoothing Parameter in Nonparametric Problems', *Journal of Multivariate Analysis*, 32, 177–203.

Hall, P., and Simar, L. (2002), 'Estimating a Changepoint, Boundary, or Frontier in the Presence of Observation Error', *Journal of the American Statistical Association*, 97, 523–534.
Hall, P., and Welsh, A.H. (1985), 'Adaptive Estimates of Parameters of Regular Variation', *The Annals of Statistics*, 13, 331–341.
Henry, I.J. (2009), 'A Harmonic Moment Tail Index Estimator', *Journal of Statistical Theory and Applications*, 8, 141–162.
Kim, M., and Kokoszka, P. (2020), 'Consistency of the Hill Estimator for Time Series Observed with Measurement Errors', *Journal of Time Series Analysis*, 41, 421–435.
Kneip, A., Simar, L., and Keilegom, I.V. (2015), 'Frontier Estimation in the Presence of Measurement Error with Unknown Variance', *Journal of Econometrics*, 184, 379–393.
Kokoszka, P., Nguyen, H., Wang, H., and Yang, L. (2020), 'Statistical and Probabilistic Analysis of Interarrival and Waiting Times of Internet2 Anomalies', *Statistical Methods & Applications*, 29, 727–744.
Kulik, R., and Soulier, P. (2011), 'The Tail Empirical Process for Long Memory Stochastic Volatility Sequences', *Stochastic Processes and their Applications*, 121, 109–134.
Leng, X., Peng, L., Zhou, C., and Wang, X. (2018), 'Endpoint Estimation for Observations with Normal Measurement Errors', *Extremes*, 21, 1–26.
Ma, S., Yan, J., and Zhang, X (2022), 'Extreme Value Modeling with Generalized Pareto Distributions for Rounded Data', Technical Report, University of Connecticut, Storrs, CT, April. Available at https://doi.org/10.1002/essoar.10511093.1.
Markovich, N. (2008), *Nonparametric Analysis of Univariate Heavy-Tailed Data: Research and Practice*, Chichester: John Wiley & Sons.
Matsui, M., Mikosch, T., and Tafakori, L. (2013), 'Estimation of the Tail Index for Lattice-Valued Sequences', *Extremes*, 16, 429–455.
Neyman, J. (1937), 'Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability', *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 236, 333–380.
Nicholson, J., Kokoszka, P., Lund, R., Kiessler, P., and Sharp, J. (2021), 'Renewal Model for Anomalous Traffic in Internet2 Links', *Statistical Modelling*, 22, 430–456.
Paulauskas, V., and Vaičiulis, M. (2013), 'On an Improvement of Hill and some Other Estimators', *Lithuanian Mathematical Journal*, 53, 336–355.
Resnick, S.I. (2007), *Heavy-Tail Phenomena*, New York: Springer.
Resnick, S.I., and Stărică, C. (1997a), 'Asymptotic Behavior of Hill's Estimator for Autoregressive Data', *Communications in Statistics. Stochastic Models*, 13, 703–721.
Resnick, S.I., and Stărică, C. (1997b), 'Smoothing the Hill Estimator', *Advances in Applied Probability*, 29, 271–293.