JOURNAL OF TIME SERIES ANALYSIS
J. Time Ser. Anal. **41**: 421–435 (2020)
Published online 01 December 2019 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1111/jtsa.12515

ORIGINAL ARTICLE

CONSISTENCY OF THE HILL ESTIMATOR FOR TIME SERIES OBSERVED WITH MEASUREMENT ERRORS

MIHYUN KIM[a] AND PIOTR KOKOSZKA[a*]
[a] *Department of Statistics, Colorado State University, Fort Collins, CO 80523, USA*

We investigate the asymptotic and finite sample behavior of the Hill estimator applied to time series contaminated by measurement or other errors. We show that for all discrete time models used in practice, whose non-contaminated marginal distributions are regularly varying, the Hill estimator is consistent. Essentially, the only assumption on the errors is that they have lighter tails than the underlying unobservable process. The asymptotic justification however depends on the specific class of models assumed for the underlying unobservable process. We show by means of a simulation study that the asymptotic robustness of the Hill estimator is clearly manifested in finite samples. We further illustrate this robustness by a numerical study of the inter-arrival times of anomalies in a backbone internet network, the Internet2 in the United States; the anomalies arrival times are computed with a roundoff error.

Received 11 June 2019; Accepted 30 October 2019

Keywords: Hill estimator; measurement error; regular variation

MOS subject classifications: 62F12, 62M10, 60G70.

1. INTRODUCTION

Our objective is to establish the consistency of the Hill estimator applied to heavy-tailed time series observed with measurement errors, and to explore the impact of the errors in finite samples. Heavy-tailed time series commonly occur in fields such as finance, insurance, hydrology, and computer network traffic. The theory of regular variation provides a suitable mathematical framework. Suppose $X_1, \dots, X_n$ is a realization of a strictly stationary time series with one-dimensional distribution function F_X, which has a regularly varying tail with index $\alpha > 0$, that is $P(X_i > x)$ behaves roughly like $x^{-\alpha}$, for large x. An estimate of α is essential for further inference related to extreme behavior of the time series. Risk measures, like the VaR or the expected shortfall, require an estimate $\hat{\alpha}$. The joint dependence structure is usually estimated by normalizing the data to the standard Fréchet distribution with $\alpha = 1$, which requires some estimate $\hat{\alpha}$. Many more applications are discussed in the monographs cited in the next paragraph. A well known and commonly used estimator of the index α is the Hill estimator, whose definition is recalled in Section 2. It is often used after an examination of the Hill plot, which is also a tool for detecting the presence of heavy tails. This article studies the Hill estimator in situations in which the data are contaminated by measurement errors.

The Hill estimator is studied in practically all monographs on extreme value theory, see for example Embrechts *et al.* (1997), Beirlant *et al.* (2006), Resnick (2007), and de Haan and Ferreira (2006). Its consistency for samples of i.i.d. random variables was first proven by Mason (1982). Consistency of the Hill estimator has been established beyond the i.i.d. setting. Hsing (1991) derives a general approach to establishing its consistency for stationary time series satisfying a certain mixing condition. Rootzén *et al.* (1990) and Rootzén (1995) also consider mixing conditions. Other extensions are obtained by Resnick and Stărică (1995), Davis and Resnick (1996), and Resnick and Stărică (1998), who show the consistency of the Hill estimator for time series using a tail empirical random

* Correspondence to: Piotr Kokoszka, Colorado State University, Fort Collins, CO 80523, USA. Email: Piotr.Kokoszka@colostate.edu

measure proposed by Mason (1988). Recently, Wang and Resnick (2019) have proven the consistency of the Hill estimator for network data in a linear preferential attachment model.

In many applications, we do not observe $X_1, \ldots, X_n$ directly. Instead, the data are measured with noise, measurement or roundoff error. In other words, we observe $Y_i = X_i + \varepsilon_i,\ i = 1, 2, \ldots, n$, where $\{\varepsilon_i\}$ is an error process. The question is whether the Hill estimator computed from the Y_i will be still consistent for α under suitable assumptions on the errors ε_i. The research presented in this article has been partially motivated by our work on modeling the stochastic behavior of internet traffic anomalies, whose arrival times are available only with a roundoff error. The database we have reports these times in 5 minutes resolution.

Putting together known results, it is fairly straightforward to establish the consistency if the X_i are i.i.d., but a more in-depth investigation is needed when they follow a stochastic process model with a complex dependence structure. We investigate this question in the context of models considered by Resnick and Stărică (1998). These include infinite moving averages with heavy-tailed innovations, bilinear processes driven by heavy-tailed noise variables, solutions of stochastic difference equations, the ARCH process of Engle (1982), and interarrival times of heavy-tailed hidden Markov chains. The models considered by Resnick and Stărică (1998) thus cover practically all known stochastic processes whose marginal distributions are regularly varying. Finite sample properties are investigated by means of a simulation study based on these models and by an application to the interarrival times of internet traffic anomalies. The main general conclusions of our research are as follows. (1) Asymptotically, the Hill estimator is robust to relatively large errors. (2) This robustness is confirmed in finite samples. (3) Five minute resolution is sufficient to estimate the tail index of the interarrival times of the anomalies we study.

Consistency of the Hill estimator based on data observed with measurement error has not been studied, but there has been considerable interest in a related problem, estimation of the end-point of a distribution function in the presence of additive observation errors, see Hall and Simar (2002), Goldenshluger and Tsybakov (2004), Kneip *et al.* (2015), and Leng *et al.* (2018). They all consider Gaussian measurement errors. We, however, do not place this restriction on the errors. We assume a broader class of error distributions. Intuitively, we can relax the assumptions on the measurement errors because heavy-tailed X_i are 'much larger' random variables than those with a finite end-point.

In Section 2, we introduce notation and assumptions. Our framework and main results are presented in Section 3. Finite sample performance of the Hill estimator in the presence of errors is investigated in Section 4. In Section 5, we present an application to the interarrival times of internet traffic anomalies. The proofs are developed in Appendix B, preceded by some preparation in Appendix A. Both appendices are placed in Supporting Information.

2. NOTATION AND ASSUMPTIONS

We start by introducing some notation, generally following Resnick (1987). Recall that $X_1, \ldots, X_n$ are non-negative random variables with common distribution F_X, which has regularly varying tail probabilities:

$$\bar{F}_X = 1 - F_X = P(X > \cdot) \in RV_{-\alpha},\ \alpha > 0. \tag{2.1}$$

We denote by X a generic random variable with the same distribution as each X_i. A function $U : \mathbb{R}_+ \to \mathbb{R}_+$ is regularly varying with index $\alpha > 0$, $U \in RV_{-\alpha}$, if for any $x > 0$,

$$\lim_{t\to\infty} \frac{U(tx)}{U(t)} = x^{-\alpha}.$$

For two functions $U, V : \mathbb{R}_+ \to \mathbb{R}_+$, we write $U(x) \sim V(x)$ if $U(x)/V(x) \to 1$, as $x \to \infty$, and $U(x) = o(V(x))$ if $U(x)/V(x) \to 0$, as $x \to \infty$.

The Hill estimator for the X_i is defined as

$$H_{k,n} := \frac{1}{k} \sum_{i=1}^{k-1} \log \frac{X_{(i)}}{X_{(k)}}, \tag{2.2}$$

DOI: 10.1111/jtsa.12515

with the convention that $X_{(1)}$ is the largest order statistic. We use definition (2.2) rather than the commonly used, asymptotically equivalent, definition with the $1/k$ replaced by $1/(k-1)$ because it leads to visually shorter formulas in the proofs. The consistency of the Hill estimator has been studied as the number of upper order statistics, k, tends to infinity with the sample size n, in such a way that $k/n \to 0$, that is

$$n \to \infty, \ k \to \infty, \ \frac{k}{n} \to 0. \tag{2.3}$$

We assume throughout the article that condition (2.3) holds.

We consider the Hill estimator based on observations contaminated by measurement or other errors whose source is difficult to quantify. We thus assume that we observe $Y_i = X_i + \varepsilon_i$, $1 \le i \le n$, where $\{\varepsilon_i\}$ are i.i.d. random errors following F_ε, and independent of the $\{X_i\}$. Then, the Hill estimator for the observations Y_i is defined as

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

In our context, $\widehat{H}_{k,n}$ is the Hill estimator that can be actually used since what we observe are the Y_i. Resnick and Stărică (1998) show that the Hill estimator based on the X_i, $H_{k,n}$, is consistent for the tail index of $\bar{F}_X$, when it is applied to certain classes of heavy-tailed stationary processes. In our context, the X_i are unobservable. We want to establish conditions on F_ε under which $\widehat{H}_{k,n}$ is consistent for the tail index of $\bar{F}_X$. We solve this problem for all classes of the X_i considered by Resnick and Stărică (1998).

The approach of Resnick and Stărică (1998) is based on the weak convergence to the measure ν on $(0, \infty]$, satisfying $\int_1^\infty \log(u)\nu(\mathrm{d}u) < \infty$. One example of the measure ν is ν_α, defined by $\nu_\alpha(x, \infty] = x^{-\alpha}$, $x > 0$. Our approach involves tail empirical random measures on $(0, \infty]$, based on the X_i, Y_i, and their weak convergence to the measure ν in $M_+(0, \infty]$, the space of Radon measures on $(0, \infty]$. We study the limit relations

$$\frac{1}{k} \sum_{i=1}^{n} I_{X_i/b(n/k)} \Rightarrow \nu, \quad \frac{1}{k} \sum_{i=1}^{n} I_{Y_i/b(n/k)} \Rightarrow \nu, \tag{2.4}$$

where $b(\cdot)$ is the quantile function, defined by

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

We investigate when the first convergence in (2.4) implies the second one. We use $\Rightarrow$ to denote weak convergence of random measures and $\stackrel{v}{\to}$ to denote vague convergence in $M_+(0, \infty]$, see Resnick (2007).

We now state assumptions on the unobservable random variables X_i. We consider several conditions. We first assume that the unobservable variables are independent and have a common, regularly varying tail distribution. We then relax this assumption by considering three classes introduced by Resnick and Stărică (1998). We first assume that the X_i follow a heavy-tailed stationary process which can be approximated by sequences of m-dependent random variables, and the m-dependent sequences carry enough information on the tail behavior of the original process. Then, we consider random coefficient autoregressive model. The final class consists of heavy-tailed hidden semi-Markov models.

Assumption 2.1. The X_i are non-negative, independent random variables with common one-dimensional distribution F_X, which has regularly varying tail probabilities, that is (2.1) holds.

Assumption 2.2. The X_i form a stationary sequence, which can be approximated by stationary m-dependent sequences $\{X_i^{(m)}\}$ as follows. There exist Radon measures $\nu^{(m)}$, ν on $(0, \infty]$ with $\int_1^\infty \log(u)\nu(\mathrm{d}u) < \infty$ and $\nu^{(m)} \stackrel{v}{\to} \nu$, as $m \to \infty$. The X_i, $X_i^{(m)}$, and the ν, $\nu^{(m)}$ satisfy the following relations.

DOI: 10.1111/jtsa.12515

(a) For any fixed $m \geq 1$ (under (2.3)),

$$\frac{n}{k} P\left(\frac{X_i^{(m)}}{b(n/k)} \in \cdot \right) \xrightarrow{v} \nu^{(m)}.$$

(b) For any $\tau > 0$,

$$\lim_{m \to \infty} \limsup_{n \to \infty} \frac{n}{k} P\left(\left| \frac{X_i^{(m)}}{b(n/k)} - \frac{X_i}{b(n/k)} \right| > \tau \right) = 0.$$

(c) For each $m \geq 1$, the function $y \mapsto \nu^{(m)}(y, \infty]$ is right-continuous.

Condition (c) is not assumed by Resnick and Stărică (1998). We need it to deal with the impact of the measurement errors. This condition is however not restrictive in practice because in all examples considered by Resnick and Stărică (1998), the functions $y \mapsto \nu^{(m)}(y, \infty]$ are continuous.

Assumption 2.3. The X_i form a stationary sequence, which satisfies the stochastic autoregressive equation

$$X_i = A_i X_{i-1} + B_i, \ -\infty < i < \infty,$$

where $\{(A_i, B_i), \ -\infty < i < \infty\}$ are i.i.d. $\mathbb{R}_+^2$-valued random pairs satisfying the following conditions. There exists $\alpha > 0$ with

$$EA_0^\alpha = 1, \ \ EA_0^\alpha \log^+ A_0 < \infty, \ \ 0 < EB_0^\alpha < \infty,$$

where $\log^+ x = \log x \vee 0$, $B_0/(1 - A_0)$ is non-degenerate, and the conditional distribution of $\log A_0$ given $A_0 \neq 0$ is non-lattice.

The conditions imposed on (A_i, B_i) ensure that the X_i are regularly varying, see Lemma A.5 (i).

The final class we consider consists of hidden semi-Markov Models. These models generalize the commonly used hidden Markov models, and have recently found application in biology, computer science, operations research and meteorology, see for example Barbu and Limnios (2008) and Yu (2010). The heavy-tailed hidden Markov has one or more states following heavy-tailed distributions. We first state its building blocks and then state the assumption. Let $\{J_n, n \geq 0\}$ be an ergodic, m-state Markov chain on the state space $\{1, 2, \ldots, m\}$ with the stationary distribution $\tilde{\pi} = (\pi_1, \ldots, \pi_m)$, and $P = \{p_{ij}, 1 \leq i, j \leq m\}$ be the transition probability matrix of the chain. Suppose $\{D_n^{(r)}, n \geq 0\}$, $r = 1, 2, \ldots, m$, are i.i.d. holding time random variables with common distributions $\{q_n^{(r)}, n \geq 0\}$, for each r. Define $\{V_i, i \geq 0\}$ by

$$V_i = \sum_{n=0}^{\infty} J_n 1_{[\sum_{l=0}^{n-1} D_l^{(j_l)} \leq i \leq \sum_{l=0}^{n} D_l^{(j_l)}]},$$

and define for $i \geq 0$,

$$X_i = F_{V_i}^{\leftarrow}(U_i), \tag{2.5}$$

where the U_i are i.i.d. uniform random variables with support $[0, 1]$, and $F_1, \ldots, F_m$ are distributions on $\mathbb{R}_+$. The $\{U_i, i \geq 0\}, \{J_n, n \geq 0\}, \{D_n^{(r)}, n \geq 0, 1 \leq r \leq m\}$ are all independent. The X_i can be thought of as interarrivals which are generated from distribution F_r when $V_i = r$.

Assumption 2.4. The X_i form a sequence satisfying (2.5) with

$$ED_n^{(r)} < \infty, \ \ r = 1, \ldots, m,$$

DOI: 10.1111/jtsa.12515

and

$$\bar{F}_1(\cdot) \in RV_{-\alpha} \text{ and } \lim_{x\to\infty} \frac{\bar{F}_j(x)}{\bar{F}_1(x)} = 0, \quad j = 2, \ldots, m. \tag{2.6}$$

Under Assumption 2.4, we define $b(\cdot)$ by $\bar{F}_1(b(t)) = t^{-1}$.

We next state an assumption on the tail distribution $\bar{F}_\varepsilon$, which says that the measurement error ε has a lighter tail than X. This assumption is reasonable as measurement errors are thought to be small relative to the quantity being measured.

Assumption 2.5. The ε_i are i.i.d. random errors with a common tail distribution $\bar{F}_\varepsilon$, which has an asymptotic tail property

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

The sequence $\{\varepsilon_i\}$ is independent of the sequence $\{X_i\}$ and of the approximating sequences $\{X_i^{(m)}\}$ in Assumption 2.2.

The order statistics used to compute the Hill estimator must be positive. In the following, all statements are tacitly assumed to hold conditional on the event $\{Y_{(k)} > 0\}$.

3. MAIN RESULTS

The underlying idea of our argument is that to get the consistency of the Hill estimator computed from error contaminated observations, it is enough to show that

$$\frac{n}{k}P\Big(\frac{Y_i}{b(n/k)} \in \cdot\Big) \xrightarrow{v} \nu \quad \text{and} \quad \frac{1}{k}\sum_{i=1}^{n} 1_{Y_i/b(n/k)} \Rightarrow \nu$$

in $M_+(0,\infty]$, where ν is the measure to which $\frac{1}{k}\sum_{i=1}^{n} 1_{X_i/b(n/k)}$ weakly converges. One can then obtain

$$\hat{H}_{k,n} \xrightarrow{P} \int_1^\infty \log(u)\nu(\mathrm{d}u), \tag{3.1}$$

by Proposition 2.4 of Resnick and Stărică (1995). If $\nu = \nu_\alpha$, defined in Section 2, (3.1) leads to

$$\hat{H}_{k,n} \xrightarrow{P} \frac{1}{\alpha}. \tag{3.2}$$

We start with the i.i.d. case. We show that $Y = X + \varepsilon$ has regularly varying tail probabilities with the same index as $\bar{F}_X$, that is $\bar{F}_Y \in RV_{-\alpha}$. This approach allows us to conclude consistency for *any* estimator of α, provided it is consistent based on the X_i. For the *Hill* estimator, regular variation of the underlying tail distribution $\bar{F}_X$ is actually equivalent to the consistency of the estimator based on the Y_i. These results are presented respectively in parts (a) and (b) of Theorem 3.1, for which Proposition 3.1 is a preparation.

Proposition 3.1. Denote $Y = X + \varepsilon$, and let $\bar{F}_Y$ be the tail distribution of Y. Suppose that $P(X > \cdot) \in RV_{-\alpha}$, $P(|\varepsilon| > x) = o(P(X > x))$, and ε is independent of X. Then,

$$\bar{F}_Y \in RV_{-\alpha}.$$

DOI: 10.1111/jtsa.12515

Theorem 3.1. (a) Under Assumptions 2.1 and 2.5, any estimator of α computed from the Y_i is consistent, if its counterpart computed from the unobservable X_i is consistent. (b) For the Hill estimator, under Assumption 2.5, the X_i satisfy Assumption 2.1 if and only if (3.2) holds.

We now turn to dependent X_i that follow one of the assumptions specified in Section 2. The contaminated variables $X_i + \epsilon_i$ need not satisfy these assumptions, and so a careful investigation is required.

We first consider the stationary process $\{X_i\}$ and its approximating m-dependent processes $\{X_i^{(m)}\}$ satisfying Assumption 2.2. Set

$$Y_i = X_i + \epsilon_i, \quad Y_i^{(m)} = X_i^{(m)} + \epsilon_i. \tag{3.3}$$

Theorem 3.2. If the unobservable sequences $\{X_i\}$ and $\{X_i^{(m)}\}$ satisfy Assumption 2.2 and if Assumption 2.5 holds, then the sequences $\{Y_i\}$ and $\{Y_i^{(m)}\}$ defined by (3.3) satisfy Assumption 2.2 as well.

Resnick and Stărică (1998) provide three examples of processes satisfying Assumption 2.2.

(a) Infinite-order moving averages of heavy-tail innovations defined by

$$X_i = \sum_{j=0}^{\infty} c_j Z_{i-j}, \ -\infty < i < \infty,$$

where the Z_i are i.i.d. non-negative random errors with a regularly varying tail distribution,

$$P(Z_i > \cdot) \in RV_{-\alpha}, \ \alpha > 0, \tag{3.4}$$

and the c_j contain at least one positive number, and satisfy $\sum_{j=0}^{\infty} |c_j|^{\delta} < \infty$, for some $0 < \delta < \alpha \wedge 1$. This model was recently studied by Bartlett and McCormick (2013).

(b) A simple bilinear model driven by heavy-tail innovations defined by

$$X_i = cX_{i-1}Z_{i-1} + Z_i, \ -\infty < i < \infty,$$

where $c > 0$ and the Z_i are i.i.d. non-negative random errors satisfying (3.4) and $c^{\alpha/2}EZ_1^{\alpha/2} < 1$.

(c) Solutions of stochastic equations of the form

$$X_i = A_iX_{i-1} + Z_i, \ -\infty < i < \infty,$$

where the Z_i are i.i.d. non-negative random errors satisfying (3.4) and $\{(A_i, Z_i) \in \mathbb{R}_+^2, -\infty < i < \infty\}$ are i.i.d. random pairs with $EA_0^{\alpha} < 1$, $EA_0^{\beta} < \infty$, for some $0 < \alpha < \beta$.

By Corollary 3.1 of Resnick and Stărică (1998), processes (a), (b), and (c) satisfy Assumption 2.2, and for process (b), α is replaced by $\alpha/2$. We thus obtain Corollary 3.1 to Theorem 3.2.

Corollary 3.1. Convergence (3.1) holds under Assumptions 2.2 and 2.5, that is (3.2) holds for the processes (a) or (c), and for the process (b), $\widehat{H}_{k,n} \xrightarrow{P} 2/\alpha$.

We next assume that the unobservable stationary process $\{X_i\}$ satisfies Assumption 2.3. In this case, it cannot be claimed that the contaminated process also satisfies Assumption 2.3. For example, if the X_i follow an ARCH model, then $X_i + \epsilon_i$ will not follow this model.

Theorem 3.3. Relations (3.1) and (3.2) hold under Assumptions 2.3 and 2.5.

DOI: 10.1111/jtsa.12515

The ARCH process introduced by Engle (1982) is defined by

$$X_i = N_i(\beta + \lambda X_{i-1}^2)^{1/2}, \ -\infty < i < \infty, \tag{3.5}$$

where the N_i are i.i.d. N(0,1) random variables, $\beta > 0$, and $\lambda > 0$. We assume $0 < \lambda < 2e^E$, where $E = 0.5772\ldots$ is Euler's constant, to guarantee the existence of α stated in Assumption 2.3, see Lemma 8.4.6 of Embrechts *et al.* (1997). The process $\{X_i^2\}$ therefore satisfies Assumption 2.3 with $A_i = \lambda N_i^2$ and $B_i = \beta N_i^2$. We then obtain Corollary 3.2 to Theorem 3.3.

Corollary 3.2. Relation (3.2) holds for the ARCH(1) process, under Assumption 2.5, provided $\beta > 0$ and $0 < \lambda < 2e^E$.

We finally study the consistency of the Hill estimator for interarrival times generated by a heavy-tailed hidden Markov model, and which are observed with measurement errors. We assume that the process $\{X_i\}$ satisfies Assumption 2.4, under which Resnick and Stărică (1998) show that $H_{k,n} \overset{P}{\to} 1/\alpha$. We consider $\widehat{H}_{k,n}$ based on $Y_i = X_i + \varepsilon_i$.

Theorem 3.4. Convergence (3.2) holds under Assumptions 2.4 and 2.5.

4. IMPACT OF MEASUREMENT ERRORS IN FINITE SAMPLES

We report the results of simulation studies of the Hill estimator applied to various processes contaminated by additive errors. We investigate the impact of these errors, especially how large they can be compared to be tolerated in practice.

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. We use four models for the X_i, those considered in Section 2.

- Model 1. The X_i are **i.i.d.** random variables, which follow a Pareto distribution with $\alpha = 2$, $P(X_i > x) = x^{-2}, \ x > 1$.
- Model 2. The X_i form the **AR(2)** process $X_i = 1.3X_{i-1} - 0.7X_{i-2} + Z_i$, where the Z_i follow a Pareto distribution with $\alpha = 2$, $P(Z_i > z) = z^{-2}, \ z > 1$.
- Model 3. The X_i form the simple **bilinear** model driven by heavy-tail innovations defined by $X_i = 0.7X_{i-1}Z_{i-1} + Z_i$, where the Z_i follow a Pareto distribution with $\alpha = 4$, $P(Z_i > z) = z^{-4}, \ z > 1$.
- Model 4. The X_i follow the **ARCH** process $X_i = N_i(1 + 0.5773X_{i-1}^2)^{1/2}$, where the N_i are i.i.d. $N(0, 1)$ random variables.

Model 2 is causal and thus has an infinite moving average representation, which satisfies Assumption 2.2. Each X_i therefore has tail index $\alpha = 2$. Model 3 also satisfies Assumption 2.2, and each X_i has tail index $\alpha/2 = 2$. Model 4 with $\beta = 1$, $\lambda = 0.5773$ satisfies Assumption 2.3, and X_i^2 has tail index α which satisfies $E(0.5773N_0^2)^\alpha = 1$. We get a numerical solution for the equation, $\alpha \approx 2$, since the equation cannot be solved explicitly. Thus, in all four models, the true value of the tail index of the X_i (X_i^2 for Model 4) is 2.

The ε_i are drawn from a normal distribution with mean 0 and standard deviation σ, a scaled t-distribution with 4 degrees of freedom (scaled t_4), and a generalized Pareto distribution (GPD), $P(|\varepsilon| > z) = (1 + \xi(z - \mu)/\sigma)^{-1/\xi}$, with location $\mu = 0$, shape $\xi = 1/4$, and scale σ_{GPD}. The scale parameters for each error distribution vary. They can be fixed or determined by the ratio of the standard deviation of error distribution (error SD) to the standard deviation of underlying process (model SD). For example, if we consider the ratio of 0.1 for Model 1 whose standard deviation is 2.88, then the corresponding scale parameter is 0.288 for the normal distribution, 0.204 for the scaled t_4, and 0.125 for the GPD. All distributions of the measurement error have a lighter tail than the X_i (X_i^2 for Model 4), so the tail distributions satisfy Assumption 2.5.

DOI: 10.1111/jtsa.12515

Table I. Empirical bias and standard error of $\hat{\alpha}$ of the Hill estimator applied to various models with additive errors following $N(0, \sigma^2)$, t_4, or GPD, with fixed error SD

		No error	$N(0, \sigma^2)$					t_4	GPD
Error SD		0	0.1	0.2	0.3	0.4	0.5	1.41	0.47
Model 1	Bias	0.02	0.05	0.05	0.07	0.08	0.11	0.31	0.07
SD = 2.88	(SE)	(0.14)	(0.11)	(0.11)	(0.11)	(0.11)	(0.12)	(0.17)	(0.24)
Model 2	Bias	0.42	0.42	0.42	0.43	0.43	0.43	0.49	0.43
SD = 6.24	(SE)	(0.33)	(0.33)	(0.33)	(0.33)	(0.33)	(0.33)	(0.34)	(0.34)
Model 3	Bias	0.23	0.23	0.23	0.22	0.23	0.23	0.23	0.23
SD = 31.5	(SE)	(0.54)	(0.54)	(0.54)	(0.53)	(0.54)	(0.53)	(0.53)	(0.54)
Model 4	Bias	−0.23	−0.22	−0.22	−0.21	−0.21	−0.21	−0.18	−0.21
SD = 7.00	(SE)	(0.21)	(0.21)	(0.21)	(0.21)	(0.22)	(0.22)	(0.22)	(0.22)

We estimate the tail index using the Hill estimator with a data-driven cut-off k, the number of upper order statistics used to compute it. We use the threshold selection method introduced by Hall (1990), which employs a bootstrap procedure to choose k that minimizes the asymptotic mean square error. This procedure is implemented by the function `hall` of the `R` package `tea`. For each model/error pair, we compute the average of the estimates over 1000 replications, and the estimated standard error based on these replications. The sample size is $n = 5000$.

Table I reports the results for fixed error SDs, Table II for fixed ratios of error SD to model SD. The error SD of 0, or the ratio 0 means that there are no errors. Model SD is calculated from the generated X_i. Table III provides information on the effects of errors on the selection of optimal k.

Tables I and II show that additive errors lead to estimates which indicate lighter tails than those indicated by the estimates computed from uncontaminated data. This can be intuitively expected because the errors in a sense 'dilute' the true heavy tails. For Models 1–3, the biases increase with the error SD. For Model 4, the bias becomes smaller in absolute value, but this cannot be interpreted that the error helps the bias to be small. Instead, this behavior is in agreement with the previous observation; the bias for uncontaminated process is negative, and it becomes less negative (lighter tail) as the error SD increases.

A rather unexpected finding is that the bias is not affected a lot even by large errors. We see from Table I that for i.i.d. X_i, Model 1, even error SD equal to half the model SD causes bias of 0.31, which is not large given the uncertainty about the selection of k. Such a level of contamination could however indicate that the data have finite variance, whereas in fact they may have infinite variance. Even more remarkable is that for dependent data with heavy-tailed marginal distributions, the errors have almost no impact on the bias and the SE of the estimator. Table II is designed to take a closer look at this finding by controlling the ratio of error SD to model SD. We first observe that the bias increases with this ratio. Second, this increase is relatively flat. Only for i.i.d. X_i, the ratio of 20 percent causes a bump in bias. For dependent X_i, such a ratio does not change the bias much compared to uncontaminated data; Models 2, 3, and 4 are surprisingly insensitive to the errors. Finally, the bias caused by the errors does not depend on the type of the error distribution. Standard errors of the estimates are basically unaffected by the errors. In some cases, the errors lead to smaller or larger estimated standard errors, but these estimates are so close that the differences are probably not statistically significant.

Table III reports on average optimal ks and their standard errors for all combinations of the underlying models and error distributions. We do not observe any clear positive or negative relationships between the average optimal k and the ratio, but the dependent data still show a considerably weaker dependence on the ratio; again, only a large increase of the ratio has a relatively large impact on the average optimal k.

Another question of interest is how the errors affect the shape of the Hill plot. The Hill estimator is *location variant*, see Section 4.2.2 of Resnick (2007). The lack of location invariance makes it sensitive to a shift in location; but it does not theoretically affect the tail index estimate. Since adding measurement errors could be thought of as

DOI: 10.1111/jtsa.12515

Table II. Empirical bias and standard error of $\hat{\alpha}$ for the fixed ratio of the error SD to model SD

Model	Error type	Error SD/model SD ratio 0	0.005	0.01	0.02	0.04	0.06	0.1	0.2
Model 1 SD = 2.88	Normal	0.02	0.04	0.05	0.05	0.05	0.05	0.06	0.12
		(0.14)	(0.13)	(0.11)	(0.11)	(0.11)	(0.11)	(0.11)	(0.12)
	Scaled t_4	0.02	0.04	0.05	0.05	0.05	0.04	0.05	0.09
		(0.14)	(0.12)	(0.11)	(0.11)	(0.11)	(0.12)	(0.12)	(0.13)
	GPD	0.02	0.05	0.05	0.05	0.04	0.04	0.04	0.07
		(0.14)	(0.12)	(0.11)	(0.11)	(0.12)	(0.12)	(0.13)	(0.14)
Model 2 SD = 6.24	Normal	0.42	0.42	0.42	0.42	0.42	0.43	0.43	0.47
		(0.33)	(0.33)	(0.33)	(0.33)	(0.33)	(0.33)	(0.33)	(0.33)
	Scaled t_4	0.42	0.42	0.42	0.42	0.43	0.43	0.43	0.47
		(0.33)	(0.33)	(0.33)	(0.33)	(0.33)	(0.33)	(0.33)	(0.33)
	GPD	0.42	0.42	0.42	0.42	0.43	0.42	0.44	0.47
		(0.33)	(0.33)	(0.33)	(0.33)	(0.34)	(0.33)	(0.34)	(0.35)
Model 3 SD = 31.5	Normal	0.23	0.23	0.23	0.23	0.22	0.23	0.24	0.26
		(0.54)	(0.54)	(0.54)	(0.54)	(0.53)	(0.53)	(0.53)	(0.53)
	Scaled t_4	0.23	0.23	0.22	0.23	0.23	0.23	0.23	0.26
		(0.54)	(0.54)	(0.54)	(0.54)	(0.54)	(0.53)	(0.53)	(0.54)
	GPD	0.23	0.23	0.23	0.23	0.23	0.23	0.23	0.27
		(0.54)	(0.54)	(0.54)	(0.54)	(0.54)	(0.53)	(0.53)	(0.55)
Model 4 SD = 7.00	Normal	−0.23	−0.22	−0.22	−0.22	−0.21	−0.21	−0.20	−0.18
		(0.21)	(0.21)	(0.21)	(0.21)	(0.21)	(0.22)	(0.22)	(0.23)
	Scaled t_4	−0.23	−0.22	−0.22	−0.22	−0.22	−0.21	−0.20	−0.18
		(0.21)	(0.21)	(0.21)	(0.21)	(0.21)	(0.22)	(0.22)	(0.23)
	GPD	−0.23	−0.22	−0.22	−0.22	−0.22	−0.21	−0.21	−0.17
		(0.21)	(0.21)	(0.21)	(0.21)	(0.21)	(0.22)	(0.22)	(0.23)

a location shift, there could be sensitivity to additive errors over some range of k, which will not show up when examining the averages. The Hill plot is a useful tool to examine this property by describing estimates as a function of the minimal order statistic k used to compute the estimates.

Figure 1 shows the Hill plots for observations generated by Model 1 with the ratio of 0.1, along with a vertical line showing the optimal k. The true tail index α is 2 for all the plots. The impact due to the additive errors, for the ratio of 0.1, turns out to be surprisingly weak as all the plots in Figure 1 look stable. We also consider the Hill plots for Model 1 observed with the same types of errors, but with a relatively large error SD corresponding to the ratio 0.2. These plots are shown in Figure 2. The shape of Hill plot is more affected by the larger error SD; this sensitivity is especially noticeable for errors with the normal and scaled t_4 distributions. The Hill plots for Models 2, 3, and 4, even without errors, do not look as stable as for i.i.d. observations, there is an upward trend. The presence of the errors changes their shape a little bit for the ratio of 0.2, but one would not say that these errors make the plots any worse.

5. APPLICATION TO INTERNET TRAFFIC ANOMALIES

We illustrate the relevance of studying the Hill estimator for error contaminated data by an application to the interarrival times of internet traffic anomalies. We first provide some background, limited in scope to conserve space, and focus on the aspects relevant to this article. More detailed network background is presented in Bandara *et al.* (2014), a paper which to some extent motivates the present research. We hope that that the analysis

DOI: 10.1111/jtsa.12515

Table III. Average optimal k (standard error) for the Hill estimator as the function of the ratio of the error SD to model SD. The sample size is 5000, and the number of replications is 1000

Model	Error type	Error SD/model SD ratio 0	0.005	0.01	0.02	0.04	0.06	0.1	0.2
Model 1 SD = 2.88	Normal	2114 (1622)	2933 (1727)	3066 (1625)	2973 (1552)	2809 (1378)	2690 (1201)	2514 (903)	2141 (537)
	Scaled t_4	2114 (1622)	2990 (1702)	3061 (1633)	2959 (1558)	2777 (1429)	2635 (1291)	2354 (1080)	1951 (735)
	GPD	2114 (1622)	3022 (1690)	3053 (1634)	2928 (1585)	2751 (1483)	2587 (1380)	2295 (1208)	1700 (892)
Model 2 SD = 6.24	Normal	1249 (827)	1250 (827)	1251 (824)	1261 (815)	1264 (811)	1280 (789)	1307 (697)	1293 (588)
	Scaled t_4	1249 (827)	1250 (826)	1254 (822)	1250 (820)	1255 (812)	1268 (787)	1225 (711)	1140 (632)
	GPD	1249 (827)	1252 (826)	1252 (826)	1248 (815)	1261 (808)	1246 (794)	1161 (725)	1041 (649)
Model 3 SD = 31.5	Normal	638 (907)	639 (907)	639 (906)	651 (908)	654 (906)	668 (901)	749 (881)	821 (827)
	Scaled t_4	638 (907)	644 (907)	642 (907)	649 (903)	652 (904)	667 (892)	703 (852)	709 (789)
	GPD	638 (907)	639 (906)	641 (906)	646 (906)	647 (902)	657 (886)	654 (831)	621 (756)
Model 4 SD = 7.00	Normal	246 (134)	242 (136)	243 (140)	241 (141)	243 (146)	248 (156)	314 (249)	510 (376)
	Scaled t_4	246 (134)	243 (136)	242 (137)	243 (141)	244 (147)	251 (157)	300 (211)	377 (267)
	GPD	246 (134)	243 (136)	244 (139)	244 (141)	246 (145)	253 (157)	281 (184)	311 (204)

presented below may guide other applications where the tail index must be estimated from error contaminated data.

Figure 3 shows the backbone internet network in the United States known as Internet2. A traffic disruption in any of the 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. An anomaly can result from a malfunction of network resources, like routers, or from malicious activity, like denial of service attacks. Bandara *et al.* (2014) developed a simple algorithm, based on the Fourier transform, which, among other characteristics, allowed them to identify the arrival time of an anomaly in any unidirectional link. They created a database of anomalies and their characteristics for 28 unidirectional links, corresponding to the 14 two-directional links shown in Figure 3, for the time period of 50 weeks, starting October 16, 2005. Due to a huge amount of data to be processed, the algorithm computes an anomaly arrival time only with the precision of five minutes. There is therefore uncertainty as to when the anomaly actually arrived, a rounding error. A key element in the analysis of anomalous traffic is to understand the distribution of the interarrival times, the time separation between the arrivals of two consecutive anomalies. This may be helpful in provisioning network resources. Bandara *et al.* (2014) perform a preliminary fitting, based on the exponential distribution. We take a closer look at this problem and place it in the context of this article.

We index the unidirectional links by integers from 1 to 28, it is not important for our analysis to which nodes they correspond. The count of anomalies detected by the algorithm of Bandara *et al.* (2014) varies from link to link, as shown in Table IV. We have examined the Hill plots and performed other diagnostic tests, and determined that it is reasonable to assume that for each link the distribution of the interarrival times is regularly varying with

DOI: 10.1111/jtsa.12515

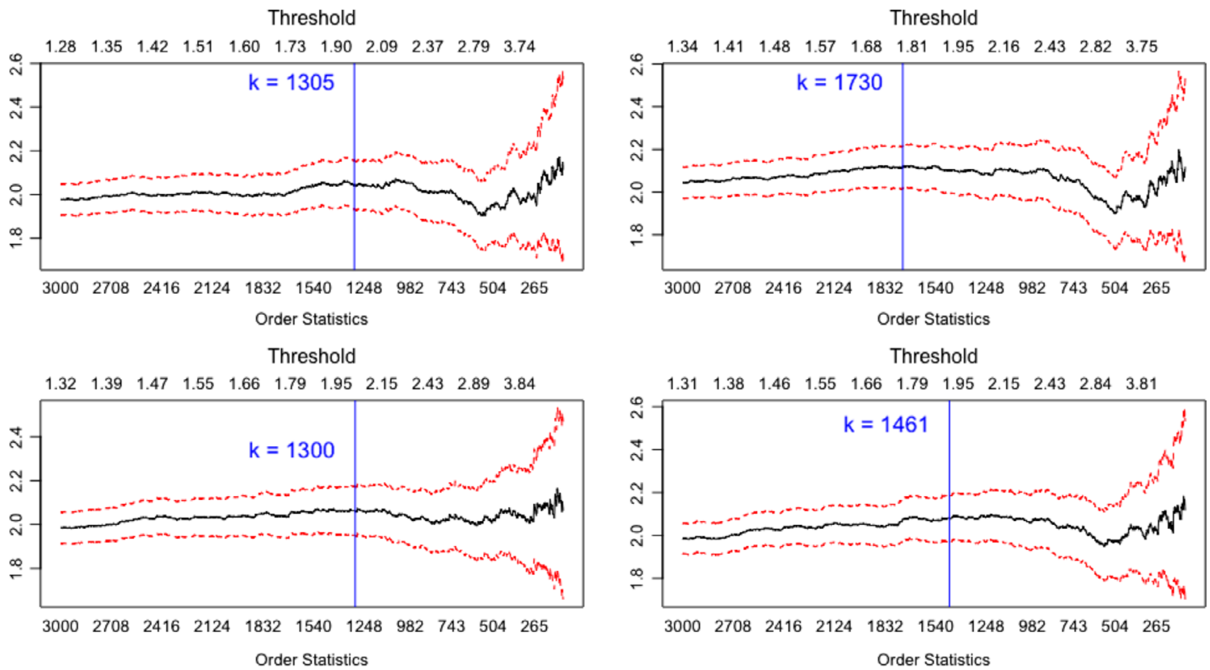

Figure 1. Hill plots with a vertical line showing the optimal k for Model 1 (a single realization) observed with no measurement errors (top left), with errors following the normal (top right), scaled t_4 (bottom left), and GPD (bottom right) with the ratio of 0.1 [Color figure can be viewed at wileyonlinelibrary.com]

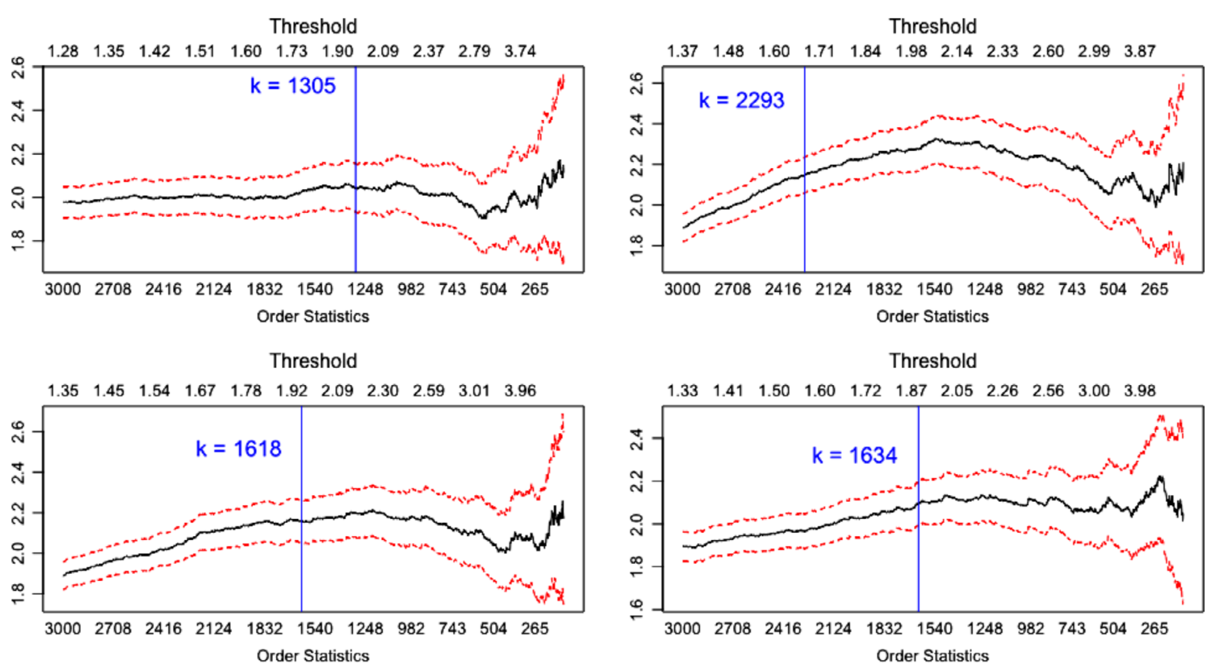

Figure 2. Hill plots with a vertical line showing the optimal k for Model 1 (a single realization) observed with no measurement errors (top left), with errors following the normal (top right), scaled t_4 (bottom left), and GPD (bottom right) with the ratio of 0.2 [Color figure can be viewed at wileyonlinelibrary.com]

Figure 3. A map showing 14 two-directional links of the Internet2 network [Color figure can be viewed at wileyonlinelibrary.com]

the tail index between 1 and 3. The values computed using the Hill estimator with the optimal k introduced in Section 4 are shown in Table IV, in the rows α_{obs}. Exploratory data analysis in Kokoszka *et al.* (2019) strongly suggests that the interarrival times form an i.i.d. sequence, so the setting of Theorem 3.1 holds.

We illustrate our analysis using the interarrival times in link 5, corresponding to anomalies traveling from Chicago to New York. In Figure 4 we display the Hill plot and the QQ plot of the log transformed data matched against exponential quantiles beyond the exceedance threshold corresponding to the optimal k. We should get approximately a line whose slope is $1/\alpha_{obs}$ if our data had a Pareto tail with index α_{obs}, see Section 4.6.4 of Resnick (2007). The QQ plot looks linear with the fit of a straight line whose slope is $1/1.53$, which tells us that it is reasonable to assume a Pareto tail with index 1.53. The same conclusion can be drawn for other links. The smallest value of α_{obs} is 1.27. It corresponds to anomalies traveling from Los Angeles to Sun Valley. The largest is 2.22, from the Indianapolis to Atlanta link.

In the context of this article, each interarrival time Y_i, computed by the algorithm, is treated as a 'true' interarrival time X_i measured with a roundoff error, that is $Y_i = X_i + \epsilon_i$. 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. Since the smallest value of Y_i in physical units is 5 minutes, we use 5 minutes as a unit lag. We therefore assume that the errors ϵ_i are uniformly distributed on $[-1, 1]$. We experimented with other beta distributions on $[-1, 1]$, the results were basically unaffected.

We perform the following numerical experiment. For each link, we generate $R = 1000$ samples of unobservable interarrival times $X_i^{(r)}$ from the observations Y_i, that is $X_i^{(r)} = Y_i - \epsilon_i^{(r)}$, where the $\epsilon_i^{(r)}$ are drawn from the uniform distribution on $[-1, 1]$, for $r = 1, \ldots, R$. We get estimates $\hat{\alpha}_r$ for each sample and then compute the average of the estimates, $\bar{\alpha}$, and their estimated standard error, σ_a, that is

$$\bar{\alpha} = \frac{1}{R}\sum_{r=1}^{R}\hat{\alpha}_r, \quad \sigma_a = \left\{\frac{1}{R}\sum_{r=1}^{R}(\hat{\alpha}_r - \bar{\alpha})^2\right\}^{1/2}.$$

The results in Table IV show that $\bar{\alpha}$ is close to α_{obs} for most links. For each link, the ratio of the error SD to the observations SD is less than 0.001, so one might expect such an outcome based on the simulations in Section 4,

DOI: 10.1111/jtsa.12515

Table IV. Results of a simulation study based on anomalous internet traffic. The tail index α_{obs} is computed from the interarrival times produced by the algorithm. The average $\bar{\alpha}$ is computed from 1000 replications of the interarrival times with errors, σ_a is the standard deviation of the 1000 estimates

Link	1	2	3	4	5	6	7
Sample size	405	247	362	454	347	345	603
α_{obs}	1.69	1.50	1.62	1.62	1.53	1.59	1.68
$\bar{\alpha}$	1.72	1.49	1.63	1.63	1.58	1.68	1.65
σ_a	0.03	0.05	0.02	0.02	0.05	0.11	0.02
Link	8	9	10	11	12	13	14
Sample size	300	387	345	382	304	476	507
α_{obs}	1.56	1.47	1.44	1.79	2.22	2.11	1.93
$\bar{\alpha}$	1.51	1.50	1.50	1.80	2.24	2.12	1.90
σ_a	0.03	0.03	0.05	0.03	0.03	0.03	0.05
Link	15	16	17	18	19	20	21
Sample size	478	319	402	388	433	493	340
α_{obs}	2.07	1.48	1.91	1.35	1.27	1.97	1.97
$\bar{\alpha}$	2.00	1.45	1.91	1.36	1.29	1.96	2.00
σ_a	0.05	0.04	0.02	0.01	0.02	0.04	0.03
Link	22	23	24	25	26	27	28
Sample size	417	597	296	258	340	348	264
α_{obs}	1.46	1.65	1.43	1.83	1.43	1.95	1.69
$\bar{\alpha}$	1.46	1.65	1.51	1.80	1.43	1.90	1.58
σ_a	0.01	0.03	0.06	0.03	0.04	0.05	0.07

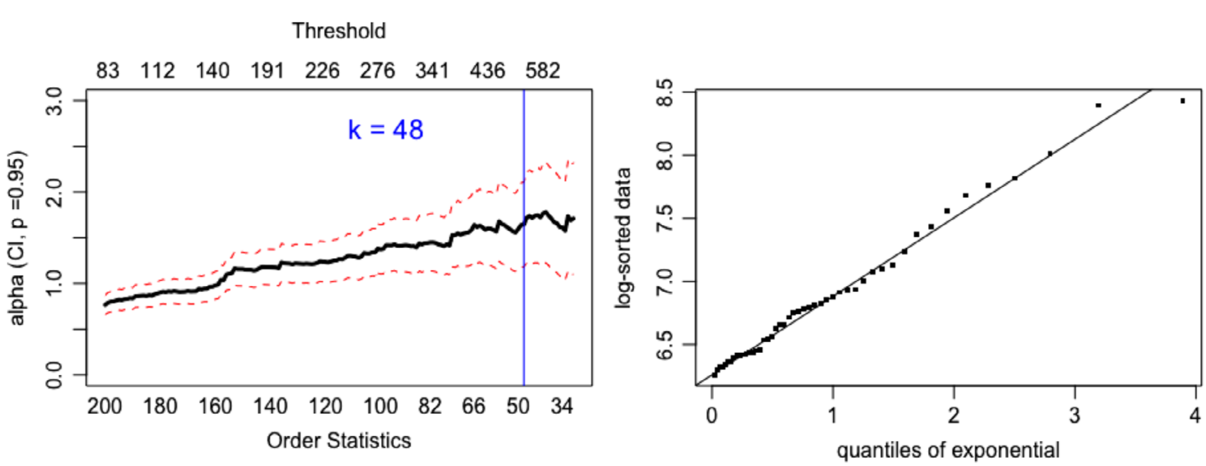

Figure 4. Hill plot (left) and QQ plot (right) for link 5 [Color figure can be viewed at wileyonlinelibrary.com]

DOI: 10.1111/jtsa.12515

but the sample sizes for these data are much smaller than 5000, so the result was not clear *a priori*. We find a couple links, 6 and 28, which have a relatively large discrepancy between α_{obs} and $\bar{\alpha}$, with a high value of σ_a. All discrepancies are however within $2\sigma_a$, so these differences are not significant. Overall, our numerical experiment shows that for the purpose of the estimation of the tail index of the anomalies interarrival times, an algorithm that identifies arrivals of anomalies with 5 minutes resolution is sufficient.

ACKNOWLEDGMENT

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

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

DATA AVAILABILITY STATEMENT

This research uses a proprietary data product derived from historical US wide internet traffic measurements. The data set and toolkit used to gather data are available for public use under Apache 2.0 license at http://www.cnrl.colostate.edu/Projects/NetworkDataAnalysis/rsDecomp.html. The data may not be uploaded to any site for further distribution.

REFERENCES

Bandara VW, Pezeshki A, Jayasumana AP. 2014. Spatiotemporal model for internet traffic anomalies. *IET Networks* **3**: 41–53.

Barbu V., Limnios N. 2008. Semi-Markov chains and hidden semi-Markov models toward applications. In *Lecture Notes in Statistics*: Springer.

Bartlett A, McCormick WP. 2013. Estimation for non-negative time series with heavy-tail innovations. *Journal of Time Series Analysis* **34**: 96–115.

Beirlant J, Goegebeur Y, Segers J, Teugels J. 2006. *Statistics of Extremes: Theory and Applications*. New York: Wiley.

Davis RA, Resnick SI. 1996. Limit theory for bilinear processes with heavy-tailed noise. *The Annals of Applied Probability* **6**: 1191–1210.

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

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

Engle R. 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* **50**: 987–1007.

Goldenshluger A, Tsybakov A. 2004. Estimating the endpoint of a distribution in the presence of additive observation errors. *Statistics & Probability Letters* **68**: 39–49.

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, Simar L. 2002. Estimating a changepoint, boundary, or frontier in the presence of observation error. *Journal of the American Statistical Association* **97**: 523–534.

Hsing T. 1991. On tail index estimation using dependent data. *The Annals of Statistics* **19**: 1547–1569.

Kneip A, Simar L, Keilegom IV. 2015. Frontier estimation in the presence of measurement error with unknown variance. *Journal of Econometrics* **184**: 379–393.

Kokoszka P., Nguyen H., Wang H., Yang L. 2019. Statistical and probabilistic analysis of interarrival and waiting times of Internet2 anomalies. *Statistical Methods & Applications*. Forthcoming.

Leng X, Peng L, Zhou C, Wang X. 2018. Endpoint estimation for observations with normal measurement errors. *Extremes* **21**: 1–26.

Mason DM. 1982. Laws of large numbers for sums of extreme values. *The Annals of Probability* **10**: 754–764.

Mason DM. 1988. A strong invariance theorem for the tail empirical process. *Annales de L'Institut Henri Poincaré* **24**: 491–506.

Resnick SI. 1987. *Extreme Values, Regular Variation, and Point Processes*. New York: Springer.

Resnick SI. 2007. *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. New York: Springer.

Resnick SI, Stărică C. 1995. Consistency of Hill's estimator for dependent data. *Journal of Applied Probability* **32**: 139–167.

Resnick SI, Stărică C. 1998. Tail index estimation for dependent data. *Annals of Applied Probability* **8**: 1156–1183.

DOI: 10.1111/jtsa.12515

Rootzén H. 1995. The tail empirical process for stationary sequences. Technical Report. Department of Mathematics, Chalmers University, Sweden.
Rootzén H., Leadbetter M., De Haan L. 1990. Tail and quantile estimation for strongly mixing stationary sequences. Technical Report. Center for Stochastic Processes, Department of Statistics, University of North Carolina.
Wang T, Resnick SI. 2019. Consistency of Hill estimators in a linear preferential attachment model. *Extremes* **22**: 1–28.
Yu SZ. 2010. Hidden semi-Markov models. *Artificial Intelligence* **174**: 215–243.