UNDERSTANDING HEAVY TAILS IN A BOUNDED WORLD OR, IS A TRUNCATED HEAVY TAIL HEAVY OR NOT?

ARIJIT CHAKRABARTY AND GENNADY SAMORODNITSKY

Abstract. We address the important question of the extent to which random variables and vectors with truncated power tails retain the characteristic features of random variables and vectors with power tails. We define two truncation regimes, soft truncation regime and hard truncation regime, and show that, in the soft truncation regime, truncated power tails behave, in important respects, as if no truncation took place. On the other hand, in the had truncation regime much of "heavy tailedness" is lost. We show how to estimate consistently the tail exponent when the tails are truncated, and suggest statistical tests to decide on whether the truncation is soft or hard. Finally, we apply our methods to two recent data sets arising from computer networks.

1. Introduction

Probability laws with power tails are ubiquitous in applications. A good fit between empirical distribution of various quantities of interest and distributions with power tails has been reported in such diverse areas as human travel (Brockmann et al. (2006)), earthquake analysis (Corral (2006)), animal science (Bartumeus et al. (2005)) and even in language (Serrano et al. (2009)). It is also true that in many situations there is a "physical" limit that prevents a quantity of interest from taking an arbitrarily large value. The File Allocation Table (FAT) used on most computer systems allows the largest file size to be 4GB (minus one byte) (Microsoft Knowledge Base Article 154997 (2007)); the greatest loss an insurance company is exposed to by an single covered event is limited by its reinsurance contract (see e.g. Mikosch (2009)). Even the number of the atoms in the universe is widely considered to be finite. It is common in practice to combine these two facts together and use a model that features power tails only in a truncated form; such models are often referred to as *truncated Lévy flights*, see e.g. Scholtz

1991 *Mathematics Subject Classification.* Primary 62G32, 62G10 Secondary 60E07.

Key words and phrases. heavy tails, truncation, regular variation, Central Limit theorem, Hill estimator, consistency .

Research partly supported by the ARO grant W911NF-07-1-0078 at Cornell University. Gennady Samorodnitsky's research was also partly supported by a Villum Kann Rasmussen Visiting Professor Grant at the University of Copenhagen and by Otto Moensted foundation grant at Danish Technological University.

and Contreras (1998), Maruyama and Murakami (2003) or Zaninetti and Ferraro (2008). At the first glance this leads to a situation where the power tails, in a sense, completely disappear. The truncation may change dramatically the behavior of the cumulative sums of observations and it always changes dramatically the behavior of the cumulative maxima of the observations. Yet it is precisely such patterns of behavior for which a model with power tails is chosen in the first place. This leads one to ask the natural question: **to what extent, if any, do phenomena well described by models with truncated power tails retain the characteristic features of power tails**?

Answering this question is not straightforward. We start by pointing out that the level of truncation is linked to the amount of observations one has at hand. This can be thought of in different ways. First of all, finiteness of the sample is sometimes taken as the source of the truncation, see e.g. Burrooughs and Tebbens (2001) or Barthelemy et al. (2008). Secondly, both the physical nature of the truncation bound and the available data can be linked to a technological level. This is particularly transparent when one models a phenomenon related to computer or communications systems; see e.g. Jelenković (1999) or Gomez et al. (2000). We describe this situation as a sequence of models, each one with truncated power tails or, in other words, as a triangular array system, which we now proceed to define formally.

Let F be a probability law on $\mathbb{R}^d$, $d \geq 1$, with the following property. There exists a sequence (b_n) with $b_n \uparrow \infty$ and a non-null Radon measure μ on $\overline{\mathbb{R}^d} \setminus \{0\}$ with $\mu\{\overline{\mathbb{R}^d} \setminus \mathbb{R}^d\} = 0$, such that

$$nF\big(b_n^{-1}\cdot\big) \xrightarrow{v} \mu(\cdot) \tag{1.1}$$

vaguely in $\overline{\mathbb{R}^d} \setminus \{0\}$. Here $\overline{\mathbb{R}^d}$ is the compactification of $\mathbb{R}^d$ obtained by adding to the latter a ball of infinite radius centered at the origin. The measure μ has necessarily a scaling property: there exists $\alpha > 0$ such that for any Borel set $B \in \mathbb{R}^d$ and $c > 0$, $\mu(cB) = c^{-\alpha}\mu(B)$. We say that the probability law F has regularly varying tails with the tail exponent α (see Resnick (1987), Hult et al. (2005)), and we view F as the law with non-truncated power tails. When studying the extent to which the central limit theorem behavior is affected by truncation (which is the main point of interest to us in the present paper) we will assume that $0 < \alpha < 2$. This restriction on the tail exponent α is precisely the one that guarantees that F is in the domain of attraction of an α-stable law; see e.g. Rvačeva (1962). Such a restriction on the values of the tail exponent will not be necessary in other parts of the paper.

For $n = 1, 2, \ldots$ (regarded both as the number of observations in the nth row of the triangular array and the number of the model) let $M_n > 0$ denote the truncation level. The nth row of the triangular array will consist of observations X_{nj}, $j = 1, \ldots, n$, which we view as generated according to the

following mechanism:

$$X_{nj} := H_j \mathbf{1}\big(\|H_j\| \le M_n\big) + \frac{H_j}{\|H_j\|}(M_n + R_j)\mathbf{1}\big(\|H_j\| > M_n\big)\,, \tag{1.2}$$

$j = 1, \dots, n$, $n = 1, 2, \dots$. Here $H_1, H_2, \dots$ are i.i.d. random vectors in $\mathbb{R}^d$ with the common law F that has regularly varying tails with a tail exponent $\alpha \in (0, 2)$, and $R_1, R_2, \dots$ are an independent of $H_1, H_2, \dots$ sequence of i.i.d. nonnegative random variables. For each $n = 1, 2, \dots$ we view the observations X_{nj}, $j = 1, \dots, n$ as having power tails that are truncated at level M_n.

We need to comment, at this point, on the role of the random variables $R_1, R_2, \dots$. One should view them as possessing light tails, even exponentially decaying tails. In many cases taking these random variables to be equal to zero with probability 1 is appropriate; in other applications exponentially fast tapering off of the tails beyond the truncation point has been observed (see e.g. Hong et al. (2008)). The reader will notice that the results of this paper hold whenever the tails of the random variables $R_1, R_2, \dots$ are only light enough, not necessarily exponentially light. We have chosen to formulate our results in this way in order to increase their generality, even though we are thinking of their role in the model (1.2) as representing the exponentially fast decaying tails.

Our approach to addressing the question "to what extent do models with truncated power tails retain the characteristic features of power tails?" lies in studying the effect of the rate of growth of the truncation level M_n on the asymptotic properties of the triangular array defined in (1.2). Specifically, we introduce the following definition. We will say that the tails in the model (1.2) are

$$\begin{array}{ll} \text{truncated softly} & \text{if } \lim_{n\to\infty} nP\big(\|H_1\| > M_n\big) = 0\,, \\ \text{truncated hard} & \text{if } \lim_{n\to\infty} nP\big(\|H_1\| > M_n\big) = \infty\,. \end{array} \tag{1.3}$$

Clearly, an intermediate regime exists as well. We will use fairly classical techniques in Section 2 below to show that, as far as the behavior of the partial sums of the truncated heavy tailed model (1.2) is concerned, observations with softly truncated tails behave like heavy tailed random variables, while observations with hard truncated tails behave like light tailed random variables. It is, however, clear that, in practice, the truncation level M_n is not observed. Therefore, we set before ourselves two tasks in this paper. The first one, is to estimate the tail exponent α based on a sample of observations with truncated power tails without knowing the truncation level or, even, if the truncation is soft or hard. We show how this can be accomplished in Section 3. The second task is to find out whether the tails in the sample are truncated softly or hard. In Section 4, where we suggest statistical procedures for testing the hypothesis of the soft (correspondingly, hard) truncation regime against the appropriate alternative. In Section 5 we

apply the statistical techniques of Section 4 to two recent data sets related to TCP connections in a large computer network.

We finish this section by pointing out that some of the issues related to models with truncated power tails have been addressed in the literature, but from different angles. The paper Asmussen and Pihlsgard (2005) discusses an application of distributions with truncated power tails in queuing, and addresses the question whether light tailed approximations or heavy approximations work better in this situation. On the other hand, a maximum likelihood estimation procedure of the tail exponent α in a parametric model of truncated power tails (specifically, the truncated Pareto distribution) is given in Aban et al. (2006). Finally, estimation of the tail exponent in randomly censored power models (where the tails are not so much truncated, as contaminated) is discussed in Beirlant et al. (2007) and Einmahl et al. (2008).

2. A Central Limit Theorem for random vectors with truncated power tails

Consider the triangular array defined in (1.2), where, as we recall, the random vectors $H_1, H_2, \ldots$ have a distribution with regularly varying tails with a tail exponent $\alpha \in (0, 2)$. This means that these random vectors (or their law F) are in the domain of attraction of some α-stable law ρ on $\mathbb{R}^d$ (see Rvačeva (1962)). That is, the partial sums $S_n^{(H)} = H_1 + \ldots + H_n$, $n = 1, 2, \ldots$, converge in law, after appropriate centering and scaling, to ρ. Defining the sums of the *truncated* observations,

$$S_n := \sum_{j=1}^{n} X_{nj},\ n = 1, 2, \ldots ,$$

we would like to know whether $\big(S_n,\, n = 1, 2, \ldots\big)$ still converge in law, after suitable centering and scaling, to ρ. If the answer is no, then we would like to know what do these sums of random vectors with truncated power tails converge to. These questions can be handled by the classical probabilistic tools, and the answer turns out to depend exclusively on the truncation regime as defined in (1.3).

2.1. **Soft truncation regime: truncated heavy tails are still heavy.**

We start with the situation where the truncation level M_n grows sufficiently fast with the sample size, so that the truncated power tails model (1.2) is in the soft truncation regime. Theorem 2.1 below shows that, in this case, the partial sums of the random vectors with truncated heavy tails converge, when properly centered and scaled, to the same α-stable limit as without truncation.

Let (c_n) and (b_n) denote, respectively, some centering and scaling sequences for the non-truncated random vectors (H_j), that is,

$$b_n^{-1} S_n^{(H)} - c_n = b_n^{-1} \sum_{j=1}^{n} H_j - c_n \Longrightarrow \rho \tag{2.1}$$

as $n \to \infty$.

Theorem 2.1. *In the soft truncation regime we have*

$$b_n^{-1} S_n - c_n \Longrightarrow \rho\,. \tag{2.2}$$

Proof. By (2.1) it is enough to show that

$$b_n^{-1} \left\| S_n - \sum_{j=1}^{n} H_j \right\| \xrightarrow{p} 0\,.$$

However, for any $\varepsilon > 0$,

$$P\left(b_n^{-1} \left\| S_n - \sum_{j=1}^{n} H_j \right\| > \varepsilon \right) \le P\Big(\|H_j\| > M_n \ \text{ for some } j = 1, \dots, n \Big)$$

$$\le nP\big(\|H_1\| > M_n \big) \to 0\,,$$

and the claim follows. □

2.2. **Hard Truncation regime: truncated heavy tails are no longer heavy.** Now we consider the situation where the truncation level M_n grows relatively slowly with the sample size, and that the truncated power tails model (1.2) is in the hard truncation regime. As we will see, in this case the partial sums of the random vectors with truncated heavy tails are no longer asymptotically α-stable but, rather, converge in law, after suitable centering and scaling, to a Gaussian limit. Therefore, at least from the point of view of the behavior of partial sums, a model with power tails that have been truncated hard does not behave anymore as a heavy tailed model.

We start with some preliminaries. Recall that, since the limiting law ρ in (2.1) is α-stable, the Lévy-Khinchine formula for its characteristic function has the form

$$\hat{\rho}(\theta) = \exp\Bigg[i\langle \theta, \gamma \rangle \tag{2.3}$$

$$+ \int_S \left(\int_0^\infty \left\{ e^{ix\langle \theta, s \rangle} - 1 - ix\langle \theta, s \rangle \mathbf{1}(x \le 1) \right\} x^{-(1+\alpha)}\, dx \right) \Gamma(ds) \Bigg]$$

for $\theta \in \mathbb{R}^d$, where $\gamma \in \mathbb{R}^d$, and Γ is a finite measure on the unit sphere in $\mathbb{R}^d$, $S := \{x \in \mathbb{R}^d : \|x\| = 1\}$, see Theorem 6.15 in Araujo and Giné (1980). The measure Γ is often referred to as spectral measure of the law ρ; see Theorem 2.3.1 in Samorodnitsky and Taqqu (1994).

Theorem 2.2. *Assume that $ER_1^2 < \infty$, and let*

$$B_n := \left[nM_n^2 P(\|H_1\| > M_n)\right]^{1/2},\ n = 1, 2, \dots .$$

Then in the hard truncation regime we have

$$B_n^{-1}(S_n - ES_n) \Longrightarrow \eta\,, \tag{2.4}$$

where η is a centered Gaussian law on $\mathbb{R}^d$ whose covariance matrix has the entries

$$\frac{2}{2-\alpha}\int_S s_i s_j\,\tilde{\Gamma}(ds),\ i, j = 1, \dots, d\,, \tag{2.5}$$

where $\tilde{\Gamma}(\cdot) := \Gamma(\cdot)/\Gamma(S)$ is the normalized spectral measure of ρ.

We start with a lemma.

Lemma 2.1. *For every continuous function $f : S \longrightarrow \mathbb{R}$,*

$$\lim_{n\to\infty} nB_n^{-2}\int_S\int_0^{M_n} f(s)r^2 P\left(\|H_1\| \in dr, \frac{H_1}{\|H_1\|} \in ds\right) = \frac{\alpha}{2-\alpha}\int_S f(s)\,\tilde{\Gamma}(ds).$$

Proof. Assumption (2.1) means that

$$\frac{P\left(\|H_1\| > r,\ \frac{H_1}{\|H_1\|} \in \cdot\right)}{P\big(\|H_1\| > r\big)} \Longrightarrow \tilde{\Gamma}(\cdot) \tag{2.6}$$

weakly on S; see e.g. Corollary 6.20 (b) of Araujo and Giné (1980). Therefore,

$$\begin{aligned}
&\int_S\int_0^{M_n} f(s)r^2 P\left(\|H_1\| \in dr, \frac{H_1}{\|H_1\|} \in ds\right)\\
&= \int_0^{M_n} 2y\left(\int_S f(s) P\left(\|H_1\| > y,\ \frac{H_1}{\|H_1\|} \in ds\right)\right) dy\\
&\qquad - M_n^2\int_S f(s) P\left(\|H_1\| > M_n,\ \frac{H_1}{\|H_1\|} \in ds\right)\\
&\sim \int_S f(s)\,\tilde{\Gamma}(ds)\left[\int_0^{M_n} 2yP\big(\|H_1\| > y\big)\,dy - M_n^2 P\big(\|H_1\| > M_n\big)\right]\\
&\sim \int_S f(s)\,\tilde{\Gamma}(ds)\left(\frac{2}{2-\alpha} - 1\right) M_n^2 P\big(\|H_1\| > M_n\big) = n^{-1}B_n^2\int_S f(s)\,\tilde{\Gamma}(ds)
\end{aligned}$$

as $n \to \infty$, where the second asymptotic equivalence follows from the Karamata theorem (see e.g. Resnick (1987)). □

Proof of Theorem 2.2. By the Cramér-Wold device it suffices to show that for every θ in $\mathbb{R}^d$,

$$B_n^{-1}\big(\langle\theta, S_n\rangle - E\langle\theta, S_n\rangle\big) \Rightarrow N\left(0, \frac{2}{2-\alpha}\int_S \langle\theta, s\rangle^2\tilde{\Gamma}(ds)\right).$$

To this end we will use the Central Limit Theorem for triangular arrays under the Lindeberg condition; see e.g. Theorem 2.4, page 345 in Gut (2005). We need to prove that

$$\lim_{n\to\infty} \frac{n}{B_n^2}\mathrm{Var}\big(\langle\theta, X_{n1}\rangle\big) = \frac{2}{2-\alpha}\int_S \langle\theta, s\rangle^2 \tilde{\Gamma}(ds) \tag{2.7}$$

and that for every $\varepsilon > 0$,

$$\frac{n}{B_n^2} E\Big(\big|\langle\theta, X_{n1}\rangle - E\big(\langle\theta, X_{n1}\rangle\big)\big|^2 \mathbf{1}\Big(\big|\langle\theta, X_{n1}\rangle - E\big(\langle\theta, X_{n1}\rangle\big)\big| > \varepsilon B_n\Big)\Big) \to 0 \tag{2.8}$$

as $n \to \infty$. In order to prove (2.7), we will show that

$$\lim_{n\to\infty} \frac{n}{B_n^2} E\big((\langle\theta, X_{n1}\rangle)^2\big) = \frac{2}{2-\alpha}\int_S \big(\langle\theta, s\rangle\big)^2 \tilde{\Gamma}(ds) \tag{2.9}$$

while

$$\lim_{n\to\infty} \frac{n^{1/2}}{B_n}\big|E\big(\langle\theta, X_{n1}\rangle\big)\big| = 0\,. \tag{2.10}$$

The former claim follows easily from Lemma 2.1 and the weak convergence (2.6) by writing

$$E\big((\langle\theta, X_{n1}\rangle)^2\big) = E\Big((\langle\theta, H_1\rangle)^2 \mathbf{1}\big(\|H_1\| \le M_n\big)\Big)$$

$$+E\left(\frac{\big(\langle\theta, H_1\rangle\big)^2}{\|H_1\|^2}(M_n + R_1)^2 \mathbf{1}\big(\|H_1\| > M_n\big)\right)$$

$$\sim n^{-1}B_n^2 \frac{\alpha}{2-\alpha}\int_S \big(\langle\theta, s\rangle\big)^2 \tilde{\Gamma}(ds) + \big(1+o(1)\big) M_n^2 E\left(\frac{\big(\langle\theta, H_1\rangle\big)^2}{\|H_1\|^2}\mathbf{1}\big(\|H_1\| > M_n\big)\right)$$

$$\sim n^{-1}B_n^2 \frac{\alpha}{2-\alpha}\int_S \big(\langle\theta, s\rangle\big)^2 \tilde{\Gamma}(ds) + M_n^2 P\big(\|H_1\| > M_n\big)\int_S \big(\langle\theta, s\rangle\big)^2 \tilde{\Gamma}(ds)$$

$$= n^{-1}B_n^2\left(\frac{\alpha}{2-\alpha} + 1\right)\int_S \big(\langle\theta, s\rangle\big)^2 \tilde{\Gamma}(ds) = n^{-1}B_n^{\alpha}\frac{\alpha}{2-\alpha}\int_S \big(\langle\theta, s\rangle\big)^2 \tilde{\Gamma}(ds)\,.$$

For (2.10) we write

$$\big|E\big(\langle\theta, X_{n1}\rangle\big)\big| \le \|\theta\|\Big[E\big(\|H_1\|\mathbf{1}(\|H_1\| \le M_n)\big) + M_n P\big(\|H_1\| > M_n\big)\Big]\,.$$

Since

$$M_n P\big(\|H_1\| > M_n\big) \ll M_n\big(P(\|H_1\| > M_n)\big)^{1/2} = n^{-1/2}B_n\,,$$

the claim (2.10) will follow once we check that

$$\lim_{n\to\infty} n^{1/2}B_n^{-1}E\left[\|H_1\|\mathbf{1}(\|H_1\| \le M_n)\right] = 0\,. \tag{2.11}$$

We give separate arguments for the cases $\alpha \leq 1$ and $\alpha > 1$.

Case 1 ($\alpha \leq 1$): Letting C be a positive constant whose value may change from line to line, by the Karamata theorem,

$$\begin{aligned} E\left[\|H_1\|\mathbf{1}(\|H_1\| \leq M_n)\right] &\leq \Big(E\Big[\|H_1\|^{3/2}\mathbf{1}(\|H_1\| \leq M_n)\Big]\Big)^{2/3} \\ &\sim CM_n\big(P(\|H_1\| > M_n)\big)^{2/3} \\ &= Cn^{-1/2}B_n\big(P(\|H_1\| > M_n)\big)^{1/6} \\ &\ll n^{-1/2}B_n\,. \end{aligned}$$

Case 2 ($1 < \alpha < 2$): Here (2.11) follows trivially from the fact that $E\left[\|H_1\|\mathbf{1}(\|H_1\| \leq M_n)\right]$ has a finite limit, while $B_n \gg n^{1/2}$ as $\alpha < 2$.

We have now proved (2.7). By (2.10), the remaining condition (2.8) will follow once we check that for every $\varepsilon > 0$,

$$\frac{n}{B_n^2}E\Big(\big|\langle\theta, X_{n1}\rangle\big|^2\mathbf{1}\big(\big|\langle\theta, X_{n1}\rangle\big| > \varepsilon B_n\big)\Big) \to 0\,.$$

This is, however, an immediate consequence of the fact that the hard truncation implies that $B_n \gg M_n$ as $n \to \infty$. □

Remark 1. We briefly address the behavior of the partial sums of the random vectors with truncated heavy tails in the intermediate regime

$$\lim_{n\to\infty} nP(\|H\| > M_n) = \delta \in (0, \infty)\,. \tag{2.12}$$

It turns out that, in this case, one can use the same centering and scaling sequences $\{c_n\}$ and $\{b_n\}$ as in the non-truncated case (2.1) (or in the soft truncation regime (2.2)), but the limit will be different. In fact,

$$b_n^{-1}S_n - c_n \Longrightarrow \rho_\delta\,, \tag{2.13}$$

where ρ_δ is an infinitely divisible law on $\mathbb{R}^d$, which is obtained by a certain truncation of the jumps of the α-stable law ρ in (2.3). Specifically,

$$\begin{aligned} \hat{\rho}_\delta(\theta) = \exp\Bigg[&i\langle\theta, \gamma_\delta\rangle \\ &+ \int_S\Bigg(\int_0^{\delta^{-1/\alpha}(\alpha^{-1}\Gamma(S))^{1/\alpha}} \Big\{e^{ix\langle\theta,s\rangle} - 1 - ix\langle\theta, s\rangle\mathbf{1}(x \leq 1)\Big\}x^{-(1+\alpha)}\,dx \\ &+\delta\Gamma(S)^{-1}\Big\{e^{i\delta^{-1/\alpha}(\alpha^{-1}\Gamma(S))^{1/\alpha}\langle\theta,s\rangle} - 1\Big\}\Bigg)\Gamma(ds)\Bigg] \end{aligned} \tag{2.14}$$

for $\theta \in \mathbb{R}^d$, where

$$\gamma_\delta = \gamma - \int_{\delta^{-1/\alpha}(\alpha^{-1}\Gamma(S))^{1/\alpha}}^{\infty} x^{-\alpha}\mathbf{1}(x \leq 1)\,dx \int_S s\,\Gamma(ds)\,.$$

We sketch the argument. Write

$$b_n^{-1}S_n - c_n = \left(b_n^{-1}\sum_{j=1}^n H_j\mathbf{1}\big(\|H_j\| \leq M_n\big) - c_n\right) \tag{2.15}$$

$$+b_n^{-1}M_n\sum_{j=1}^{n}\frac{H_j}{\|H_j\|}\mathbf{1}\big(\|H_j\|>M_n\big)+b_n^{-1}\sum_{j=1}^{n}\frac{H_j}{\|H_j\|}R_j\mathbf{1}\big(\|H_j\|>M_n\big)\,.$$

It is easy to check that the last term in the right hand side of (2.15) converges to zero in probability. Since (2.12) implies that

$$\frac{M_n}{b_n}\to\delta^{-1/\alpha}(\alpha^{-1}\Gamma(S))^{1/\alpha} \tag{2.16}$$

as $n\to\infty$, Theorem 5.9, p. 129, of Araujo and Giné (1980) implies that the first term in the right hand side of (2.15) has a weak limit whose characteristic function is given by

$$\exp\Bigg[i\langle\theta,\gamma_\delta\rangle$$

$$+\int_S\Bigg(\int_0^{\delta^{-1/\alpha}(\alpha^{-1}\Gamma(S))^{1/\alpha}}\Big\{e^{ix\langle\theta,s\rangle}-1-ix\langle\theta,s\rangle\mathbf{1}(x\le1)\Big\}x^{-(1+\alpha)}\,dx\Bigg)\Gamma(ds)\Bigg]\,.$$

Finally, it follows from (2.16) that the second term in the right hand side of (2.15) is asymptotically equivalent to

$$\delta^{-1/\alpha}(\alpha^{-1}\Gamma(S))^{1/\alpha}\sum_{j=1}^{n}\frac{H_j}{\|H_j\|}\mathbf{1}\big(\|H_j\|>M_n\big)\,,$$

and, by (2.6) and (2.12), the sum above converges weakly to the law of the Poisson sum $\sum_{j=1}^{N}Y_j$, where $Y_1,Y_2,\dots$ are i.i.d. S-valued random variables with the common law $\tilde\Gamma$, and N is an independent of them Poisson random variable with mean δ. Since the weak limits of the first and the second terms in the right hand side of (2.15) are easily seen to be independent, this shows (2.13).

3. Hill estimator for random variables with truncated power tails

Estimating the tail exponent α is one of the main statistical issues one faces when working with data for which a model with power tails is contemplated. This is a difficult statistical problem because one attempts to estimate a parameter governing the tail behavior in an otherwise nonparametric model. By necessity, any estimator one uses has to be based on a vanishing fraction of the available data. The situation is even trickier when one tries to estimate the tail exponent in a sample of observations with truncated power tails. This is the task we address in this section.

The formal setup in this section is as follows. We are given a sample $X_1,\dots,X_n$ of **one-dimensional nonnegative observations** from the model with truncated power tails, i.e. (1.2). We emphasize a slight change in notation from (1.2): whereas the latter used the notation $X_{n1},\dots,X_{nn}$ to emphasize the triangular array nature of the model, in a statistical procedure, when a single sample (*i.e.*, a particular row of the triangular array) is given, the notation $X_1,\dots,X_n$ is more natural. The discussion in Section

2 makes it intuitive that estimating the tail exponent α should be easier if the tails are truncated softly, than in the case when the tails are truncated hard. This is, indeed, the case. However, in this section we are interested in finding a procedure that permits consistent estimation of the tail exponent α regardless of the truncation regime; this is especially important because the truncation regime is never known (see, however, Section 4 below). Furthermore, in this section we do not restrict the values of the tails exponent to the interval $(0, 2)$. That is, α can take any positive value.

A number of estimators of the tail exponent of distributions with nontruncated power tails have been suggested; a thorough discussion can be found in Chapter 4 of de Haan and Ferreira (2006). One of the best known and widely used estimators is the *Hill estimator* introduced by Hill (1975). Given a sample $X_1, \ldots, X_n$. the Hill statistic is defined by

$$h_{n,k} = \frac{1}{k-1} \sum_{i=1}^{k-1} \log \frac{X_{(i)}}{X_{(k)}}\,, \tag{3.1}$$

where $X_{(1)} \geq X_{(2)} \geq \ldots \geq X_{(n)}$ are the order statistics from the sample $X_1, \ldots, X_n$, and $k = 2, \ldots, n$ is a user-determined parameter, the number of the upper order statistics to use in the estimator. The consistency result for the Hill estimator says that, if $X_1, \ldots, X_n$ are i.i.d. with regularly varying right tail with exponent $\alpha > 0$, and $k = k_n \to \infty$, $k_n/n \to 0$ as $n \to \infty$, then $h_{n,k_n} \to 1/\alpha$ in probability as $n \to \infty$; see e.g. Theorem 3.2.2 in de Haan and Ferreira (2006).

In spite of the simplicity of the statement of the consistency of the Hill estimator, selecting the number k of the upper order statistics for a given sample with nontruncated power tails remains a daunting problem; see e.g. pp. 192-193 in Embrechts et al. (1997). In the main result of this section, Theorem 3.1 below, we will see that one has to be particularly careful when using the Hill estimator on a sample with truncated power tails. Nonetheless, a consistent estimator can still be obtained.

Notice that the next theorem does not impose any conditions on the random variables $R_1, R_2, \ldots$ in the model (1.2).

Theorem 3.1. *Suppose that the number k_n of the upper order statistics satisfies*

$$nP(H > M_n) + 1 \ll k_n \ll n\,. \tag{3.2}$$

Then $h_{n,k_n} \to 1/\alpha$ in probability as $n \to \infty$.

Note that Theorem 3.1 says that, in the soft truncation regime, the Hill estimator is consistent under the same assumption, $k_n/n \to 0$, as in the nontruncated case.

Proof. For simplicity, we write k instead of k_n. An inspection of the proof of consistency of the Hill estimator in the nontruncated case in e.g. Resnick

(2007) shows that the result will follow once we check that, under the conditions of the theorem,

$$\frac{n}{k} P\left[\frac{X_{n1}}{b(n/k)} \in \cdot\right] \xrightarrow{v} \mu(\cdot) \tag{3.3}$$

vaguley in $(0,\infty]$, where μ is a measure on $(0,\infty]$ defined by

$$\mu((x,\infty]) = x^{-\alpha} \text{ for all } x > 0\,,$$

and

$$b_n = \inf\{x > 0 :\ P(H_1 > x) \le n^{-1}\},\ n = 1, 2, \ldots\,.$$

Note that (b_n) is no longer necessarily a sequence satisfying (2.1). In fact, we will use this notation several times in the sequel to denote other quantile-type functions associated with the random variable H_1.

By the hypothesis,

$$\lim_{n\to\infty} \frac{n}{k} P(H_1 > M_n) = 0$$

and, hence, $b(n/k) \ll M_n$ as $n \to \infty$. Therefore, for any $x > 0$, for n large enough,

$$\begin{aligned} P\left[\frac{X_{n1}}{b(n/k)} > x\right] &= P\big(H_1 > x b(n/k)\big) \\ &\sim \frac{k}{n} x^{-\alpha} \end{aligned}$$

where the last line follows from the hypothesis $k \ll n$ and regular variation of the tail of H_1. This shows (3.3). □

Since the truncation level M_n is not known, it is desirable to have a sample-based way of deciding on the number of upper order statistics to use in the Hill estimator. A natural (in view of the condition (3.2)) choice is to use *a random number* of upper order statistics given by

$$\hat{k}_n = \left[n \left(\frac{1}{n} \sum_{j=1}^{n} \mathbf{1}\big(X_j > \gamma \max_{i=1,\ldots,n} X_i\big) \right)^{\beta} \right], \tag{3.4}$$

where γ and β are user-specified parameters taking values in $(0,1)$ and $[\cdot]$ denotes the integer part. We will show in a separate publication that this choice of the number of upper order statistics leads to a consistent estimator of the reciprocal of the tail exponent.

4. Testing for soft and hard truncation

The first two sections of this paper provide, among other things, evidence that, in certain important respects, random variables with truncated heavy tails retain "most of the tail heaviness" if the truncation is soft, but loose "much of the tail heaviness" if the truncation is hard. Since the truncation level is not observed, how does one decide if the tails of observed data have

been truncated softly or hard? In this section we construct statistical tests for testing each of the two hypothesis against the corresponding alternative. As in Section 3 we restrict ourselves to the case of one-dimensional observations and the tail exponent α can take any positive value.

Suppose that we are given a sample $X_1, \ldots, X_n$ of one-dimensional observations from the model (1.2). As in Section 3, we do not use here the triangular array notation. Neither the precise value of the tail exponent nor the exact distribution of the random variables (R_n) in (1.2) are assumed to be known. However, we will assume that an upper bound on the tail exponent α is known.

This section is split into three subsection, describing, correspondingly, testing the hypothesis of soft truncation, testing the hypothesis of hard truncation, and testing a slightly stronger version of the latter.

4.1. **Testing the hypothesis of soft truncation.** We consider the following problem of testing a null hypothesis against a simple alternative:

$$\left.\begin{array}{ll} H_0: & P(|H_1| > M) \ll n^{-1} \quad \text{(soft truncation)} \\ H_1: & P(|H_1| > M) \gg n^{-1} \quad \text{(hard truncation)} \end{array}\right\}. \tag{4.1}$$

We assume the tail exponent α satisfies

$$\alpha < A < \infty\,, \tag{4.2}$$

i.e. an upper bound on the tail exponent is available. As a test statistic we will use

$$Z_n(A) := \frac{\sum_{i=1}^n |X_i|^A}{\max_{1\le i\le n} |X_i|^A}\,. \tag{4.3}$$

The following proposition describes the asymptotic distribution of $Z_n(A)$ under the null hypothesis and under the alternative.

Proposition 4.1. *(i) Under the hypothesis H_0 of soft truncation,*

$$Z_n(A) \Rightarrow \Gamma_1^{A/\alpha} \sum_{j=1}^{\infty} \Gamma_j^{-A/\alpha}\,, \tag{4.4}$$

where $(\Gamma_j,\, j \ge 1)$ are the arrival times of a unit rate Poisson process on $(0,\infty)$.

(ii) Assume that $ER_1^A < \infty$. Then under the hypothesis H_1 of hard truncation, $Z_n(A) \stackrel{P}{\longrightarrow} \infty$.

Proof. For part (i), we define

$$b_n = \inf\bigl\{x > 0:\, P(|H_1|^A > x) \le n^{-1}\bigr\},\ n = 1, 2, \ldots\,.$$

Note that, for any $x > 0$,

$$nP\bigl(b_n^{-1}|X_1|^A > x\bigr) \sim nP\bigl(b_n^{-1}|H_1|^A > x\bigr) \to x^{-\alpha}$$

as $n \to \infty$. It follows from Proposition 3.21 (page 154) in Resnick (1987) that we have the following weak convergence of a sequence of point processes on $(0, \infty]$:

$$N_n := \sum_{j=1}^{n} \delta_{b_n^{-1}|X_1|^A} \Rightarrow N := \sum_{j=1}^{\infty} \delta_{\Gamma_j^{-A/\alpha}} \tag{4.5}$$

as $n \to \infty$. Here δ_a is a point mass at a, and the weak convergence takes place in the space of Radon point measures on $(0, \infty]$ endowed with the topology of vague convergence; see Section 3.4 in Resnick (1987). We would like to use the continuous mapping theorem to deduce (4.4) from (4.5), but a preliminary truncation step is necessary.

For $\varepsilon > 0$ we define

$$Z_n(A; \varepsilon) := \frac{\sum_{i=1}^{n} |X_i|^A \mathbf{1}\big(b_n^{-1}|X_i|^A > \varepsilon\big)}{\max_{1 \le i \le n} |X_i|^A} \,.$$

Notice that $Z_n(A; \varepsilon) = h(N_n)$, where for a Radon point measure $\eta = \sum_j \delta_{r_j}$ on $(0, \infty]$,

$$h(\eta) = \frac{\eta\big((\varepsilon, \infty]\big)}{\max_j r_j} \,.$$

It is standard (and easy) to check that h is continuous with probability 1 at the Poisson random measure N in (4.5), so by the continuous mapping theorem,

$$Z_n(A; \varepsilon) \Rightarrow \Gamma_1^{A/\alpha} \sum_{j=1}^{\infty} \Gamma_j^{-A/\alpha} \mathbf{1}\big(\Gamma_j^{-A/\alpha} > \varepsilon\big) \,.$$

Therefore, the convergence (4.4) will follow once we check that for every $\delta > 0$,

$$\lim_{\varepsilon \to 0} \limsup_{n \to \infty} P\big(Z_n(A) - Z_n(A; \varepsilon) > \delta\big) = 0 \,. \tag{4.6}$$

To this end, notice that, for any $0 < \theta < 1$ we can select $\tau > 0$ so small that $P\big(\max_{1 \le i \le n} |X_i|^A \le \tau b_n\big) \le \theta$ for all n large enough. Then, for all n large enough,

$$\begin{aligned}
P\big(Z_n(A) - Z_n(A; \varepsilon) > \delta\big) &\le \theta + \delta^{-1} E\left(\tau^{-1} b_n^{-1} \sum_{i=1}^{n} |X_i|^A \mathbf{1}\big(b_n^{-1}|X_i|^A \le \varepsilon\big)\right) \\
&= \theta + \delta^{-1}\tau^{-1} n b_n^{-1} E\Big(|X_1|^A \mathbf{1}\big(b_n^{-1}|X_1|^A \le \varepsilon\big)\Big) \\
&= \theta + \delta^{-1}\tau^{-1} n b_n^{-1} E\Big(|H_1|^A \mathbf{1}\big(b_n^{-1}|H_1|^A \le \varepsilon\big)\Big) \\
&\sim \theta + \delta^{-1}\tau^{-1} n b_n^{-1}\Big((1 - \alpha/A)^{-1}(\varepsilon b_n) P\big(|H_1|^A > \varepsilon b_n\big)\Big) \\
&\sim \theta + \delta^{-1}\tau^{-1} n b_n^{-1}(1 - \alpha/A)^{-1}(\varepsilon b_n)\big(\varepsilon^{-\alpha/A} n^{-1}\big) \\
&= \theta + \delta^{-1}\tau^{-1}(1 - \alpha/A)^{-1}\varepsilon^{1-\alpha/A}.
\end{aligned}$$

where the second equality holds because of soft truncation, and the first asymptotic equivalence follows from the Karamata theorem. Since $A > \alpha$, we obtain (4.6) by first letting $\varepsilon \to 0$ and then $\theta \to 0$. This completes the proof of part (i).

For part (ii), we start with observing that

$$\frac{\sum_{i=1}^n |X_i|^A}{nM_n^A P(|H_1| > M_n)} \geq \frac{\sum_{i=1}^n |H_i|^A \mathbf{1}\big(M_n/2 \leq |H_i| \leq M_n\big)}{nM_n^A P(|H_1| > M_n)} \tag{4.7}$$

$$\geq (M_n/2)^A \frac{\sum_{i=1}^n \mathbf{1}\big(M_n/2 \leq |H_i| \leq M_n\big)}{nM_n^A P(|H_1| > M_n)} \sim 2^{-A}(2^\alpha - 1)$$

in probability. On the other hand, for some constant $c > 0$, by the assumption $ER_1^A < \infty$,

$$\max_{1\leq i\leq n} |X_i|^A \leq c\big(M_n^A + \max_{1\leq j\leq n} R_j^A\big) = cM_n^A + o(1)n$$

a.s. as $n \to \infty$. Since the truncation is hard, and $A > \alpha$, we see that

$$\frac{\max_{i=1,\ldots,n} |X_i|^A}{nM_n^A P(|H_1| > M_n)} \to 0 \tag{4.8}$$

a.s. as $n \to \infty$ as well. The claim of part (ii) follows from (4.7) and (4.8). □

Based on Proposition 4.1, we suggest the following test for the problem (4.1).

(4.9) reject H_0 at significance level $p \in (0,1)$ if $Z_n(A) > c_p(\alpha/A)$,

with $c_p(\theta)$ such that $P(Z(\theta) > c_p(\theta)) = p$, where for $0 < \theta < 1$,

$$Z(\theta) = \Gamma_1^{1/\theta} \sum_{j=1}^\infty \Gamma_j^{-1/\theta}. \tag{4.10}$$

The random variable $Z(\theta)$ does not seem to have one of the standard distributions, and we are not aware of any previous studies of the distribution of $Z(\theta)$. The following proposition lists some of the properties of this distribution.

Proposition 4.2. *The random variable $Z(\theta)$ is an infinitely divisible random variable. It has a density with respect to the Lebesgue measure, and the Laplace transform*

$$Ee^{-\gamma Z(\theta)} = \left(1 + \gamma e^\gamma \int_0^1 e^{-\gamma x} x^{-\theta}\, dx\right)^{-1}, \tag{4.11}$$

$\gamma > \gamma_0$, where $\gamma_0 < 0$ is the number satisfying

$$1 + \gamma_0 e^{\gamma_0} \int_0^1 e^{-\gamma_0 x} x^{-\theta}\, dx = 0.$$

Proof. For $\delta > 0$ let

$$W_\delta = \sum_{j=1}^{\infty} \big(\delta + \Gamma_j\big)^{-1/\theta}\,.$$

Then W_δ is an infinitely divisible random variable with the Laplace transform

$$Ee^{-\gamma W_\delta} = \exp\left\{-\int_0^{\delta^{-1/\theta}} \big(1 - e^{-\gamma y}\big)\theta y^{-(1+\theta)}\,dy\right\}$$

for all $\gamma \in \mathbb{R}$ because the Lévy measure of W_δ has a compact support; see Rosiński (1990) and Sato (1999). Since

$$Z(\theta) \stackrel{d}{=} 1 + T^{1/\theta} W_T$$

where T is a standard exponential random variable independent of $\big(\Gamma_j : j \geq 1\big)$, it follows that

$$\begin{aligned} Ee^{-\gamma Z(\theta)} &= \int_0^\infty e^{-t} e^{-\gamma} Ee^{-\gamma t^{1/\theta} W_t}\,dt \qquad (4.12)\\ &= e^{-\gamma}\int_0^\infty e^{-t} \exp\left\{-t\int_0^1 \big(1 - e^{-\gamma x}\big)\theta x^{-(1+\theta)}\,dx\right\} dt\\ &= e^{-\gamma}\int_0^\infty \exp\left\{-t\left[e^{-\gamma} + \gamma\int_0^1 e^{-\gamma x} x^{-\theta}\,dx\right]\right\} dt \end{aligned}$$

via integration by parts. Since the exponent under the integral is positive if and only if $\gamma > \gamma_0$, we obtain (4.11). Additionally, it follows from (4.12) that

$$Z(\theta) \stackrel{d}{=} 1 + Y(T)\,, \qquad (4.13)$$

where $\big(Y(t),\, t \geq 0\big)$ is a subordinator satisfying

$$Ee^{-\gamma Y(t)} = \exp\left\{-t\int_0^1 \big(1 - e^{-\gamma x}\big)\theta x^{-(1+\theta)}\,dx\right\}, \quad t \geq 0\,, \qquad (4.14)$$

independent of T. Since a Lévy process stopped at an independent infinitely divisible random time is, obviously, infinitely divisible, so is $Z(\theta)$. Furthermore, the characteristic function of $Y(t)$ is integrable of the real line for every $t > 0$, so each $Y(t)$ has a density, and then the same is true for any mixture of $(Y(t))$. Therefore, $Z(\theta)$ has a density. □

Even though we know, by Proposition 4.2, that the random variable $Z(\theta)$ has a density, at present we do not know ways to compute this density. One possibility to estimate the critical values $c_p(\alpha/A)$ to perform the test (4.9), is as follows. For values of α not too close to the upper bound A (or, equivalently, for the values of θ not too close to 1), it is possible to estimate the critical values by the Monte-Carlo method, by truncating the the infinite series at a sufficiently large finite number of terms. Using $N = 10^5$ number of terms in the series and generating the (truncated) random variable 10^5 times, we have estimated the following quantiles, for a range of values θ.

p \ θ	0.5	0.6	0.7
.05	4.3	5.8	8.2
.025	5.1	6.9	9.8
.01	6.2	8.4	12.1

For θ closer to 1, the rate of convergence of the truncated sum $\sum_{j=1}^{N} \Gamma_j^{-1/\theta}$ as $N \to \infty$ is very slow, and in order to obtain upper bounds on the quantiles of the random variable $Z(\theta)$ we used Proposition 4.2 as described below. Such upper bounds lead to conservative versions of the test (4.9). We use the exponential Markov inequality: for $0 < r < -\gamma_0$,

$$P(Z(\theta) \geq z) \leq e^{-rz} E e^{rZ} = e^{-rz} \left(1 - re^{-r} \int_0^1 e^{rx} x^{-\theta}\, dx\right)^{-1},$$

and estimate the integral from above by

$$\int_0^1 e^{rx} x^{-\theta} dx \leq e^{r/k} \frac{k^{\theta-1}}{1-\theta} + \frac{1}{k} \sum_{j=2}^{k} e^{rj/k} \left(\frac{j-1}{k}\right)^{-\theta},$$

$k > 1$. Using $r = .05$ and $k = 10^7$ we computed numbers $\tilde{c}_p(\theta)$ satisfying

$$P\big(Z(\theta) \geq \tilde{c}_p(\theta)\big) \leq p\,.$$

These numbers $\tilde{c}_p(\theta)$ are reported in the following table.

p \ θ	0.8	0.9	0.95
.05	65.43	73.12	127.37
.025	79.29	86.98	141.23
.01	97.62	105.31	159.56

Since we are only assuming that the tail exponent α has a known upper bound as in (4.2), but the exact value of α may be unknown, a possible way to obtain a conservative estimate of the critical value $c_p(\alpha/A)$ in (4.10) is to choose a number $A_1 > A$ and use the statistic $Z_n(A_1)$ instead of $Z_n(A)$ in (4.3). By Proposition 4.1, under the null hypothesis, the test statistic converges weakly to $Z(\alpha/A_1)$, which is stochastically smaller than $Z(A/A_1)$, and we obtain a conservative test by modifying (4.9) as follows:

(4.15) reject H_0 at significance level $p \in (0,1)$ if $Z_n(A_1) > c_p(A/A_1)$.

4.2. **Testing the hypothesis of hard truncation.** In this subsection we consider the following problem of testing a null hypothesis against a simple alternative:

$$\left.\begin{array}{lll} H_0: & P(|H_1| > M) \gg n^{-1} & \text{(hard truncation)} \\ H_1: & P(|H_1| > M) \ll n^{-1} & \text{(soft truncation)} \end{array}\right\}. \tag{4.16}$$

We still assume that an upper bound (4.2) on the tail exponent is known. For a test statistic in this case we choose a number $\gamma \in (0,1)$ and define

$$Z_n(A;\gamma) = \frac{\left(\sum_{j=1}^{[\gamma n]} (-1)^j X_j^{\langle A/2\rangle}\right)^2}{\sum_{j=[\gamma n]+1}^{n} |X_j|^A}. \tag{4.17}$$

Here $a^{\langle b\rangle} = |a|^b \mathrm{sign}(a)$ for real a, b is the signed power. The asymptotic distribution of $Z_n(A;\gamma)$ under the null hypothesis and under the alternative in (4.16) is described in Proposition 4.3 below. Recall the standard notation of $S_\alpha(\sigma,\beta,\mu)$ for (the distribution of) an α-stable random variable with the scale σ, skewness β and location μ; see Samorodnitsky and Taqqu (1994). For a symmetric α-stable random variable, $\beta = \mu = 0$. For a positive strictly α-stable random variable with $0 < \alpha < 1$, one has $\beta = 1$ and $\mu = 0$. Finally, for $0 < \alpha < 2$, let

$$C_\alpha = \begin{cases} (\Gamma(1-\alpha)\cos(\pi\alpha/2))^{-1} & \text{if } \alpha \neq 1, \\ 2/\pi & \text{if } \alpha = 1, \end{cases}$$

Proposition 4.3. *(i) Assume that $ER_1^{2A} < \infty$. Then under the hypothesis H_0 of hard truncation,*

$$Z_n(A;\gamma) \Rightarrow C_1(\gamma)\chi_1^2\,, \tag{4.18}$$

where $C_1(\gamma) = 2\gamma/(1-\gamma)$, and χ_1^2 is the standard chi-square random variable with one degree of freedom.

(ii) Under the hypothesis H_1 of soft truncation,

$$Z_n(A;\gamma) \Rightarrow C_2(A;\gamma)\frac{S_1^2}{S_2}\,, \tag{4.19}$$

where

$$C_2(A;\gamma) = \left(\frac{\gamma}{1-\gamma}\frac{C_{\alpha/A}}{C_{2\alpha/A}}\right)^{A/\alpha}\,,$$

and S_1 and S_2 are independent random variables, such that S_1 is a symmetric $2\alpha/A$-stable random variable with unit scale, and S_2 is a positive strictly α/A-stable random variable with unit scale.

Proof. The claim of part (i) will follow from the following two statements.

$$\frac{1}{(nM_n^A P(|H_1| > M_n))^{1/2}} \sum_{j=1}^{[\gamma n]} (-1)^j X_j^{\langle A/2\rangle} \Rightarrow \left(\frac{2A\gamma}{A-\alpha}\right)^{1/2} N(0,1)\,, \tag{4.20}$$

and

$$\frac{1}{nM_n^A P(|H_1| > M_n)} \sum_{j=[\gamma n]+1}^{n} |X_j|^A \to \frac{A(1-\gamma)}{A-\alpha} \tag{4.21}$$

in probability. We prove (4.21) first, and it is enough to show that

$$\frac{1}{nM_n^A P(|H_1|>M_n)} E\left(\sum_{j=[\gamma n]+1}^{n} |X_j|^A\right) \to \frac{A(1-\gamma)}{A-\alpha} \tag{4.22}$$

and

$$\frac{1}{\big(nM_n^A P(|H_1|>M_n)\big)^2} \mathrm{Var}\left(\sum_{j=[\gamma n]+1}^{n} |X_j|^A\right) \to 0\,. \tag{4.23}$$

Note that by the Karamata theorem,

$$E\left(\sum_{j=[\gamma n]+1}^{n} |X_j|^A\right) \sim (1-\gamma)n\, E\big(|X_1|^A\big)$$

$$= (1-\gamma)n\Big[E\big(|H_1|^A \mathbf{1}(|H_1|\le M_n)\big) + E(M_n+R_1)^A P(|H_1|>M_n)\Big]$$

$$\sim (1-\gamma)n\left[\frac{\alpha}{A-\alpha} M_n^A P(|H_1|>M_n) + M_n^A P(|H_1|>M_n)\right]$$

$$= \big(nM_n^A P(|H_1|>M_n)\big)\frac{A(1-\gamma)}{A-\alpha}\,,$$

proving (4.22). A similar calculation gives us

$$\mathrm{Var}\left(\sum_{j=[\gamma n]+1}^{n} |X_j|^A\right) \sim (1-\gamma)n\,\mathrm{Var}\big(|X_1|^A\big)$$

$$\le n\, E\big(|X_1|^{2A}\big) \sim \big(nM_n^{2A} P(|H_1|>M_n)\big)\frac{2A}{2A-\alpha}\,,$$

and (4.23) follows because the truncation is hard. Therefore, we have established (4.21).

In order to prove (4.20), note that the triangular array

$$\tilde{X}_{nj} := H_j^{\langle A/2\rangle}\mathbf{1}\big(|H_j|^{A/2}\le M_n^{A/2}\big) + \frac{H_j}{|H_j|}(M_n^{A/2}+R_j^{A/2})\mathbf{1}\big(|H_j|^{A/2}>M_n^{A/2}\big)\,,$$

$j=1,\dots,n$, $n=1,2,\dots$, satisfies the assumptions of Theorem 2.2 (with α replaced by $2\alpha/A$), and, therefore,

$$\frac{1}{\big(nM_n^A P(|H_1|>M_n)\big)^{1/2}}\left(\sum_{j=1}^{n}\tilde{X}_{nj} - E\Big(\sum_{j=1}^{n}\tilde{X}_{nj}\Big)\right) \Rightarrow \left(\frac{2A}{A-\alpha}\right)^{1/2} N(0,1)\,.$$

The random variables $\big(X_j^{\langle A/2\rangle}\big)$ form a somewhat different triangular array, namely

$$X_{nj}^{\langle A/2\rangle} = H_j^{\langle A/2\rangle}\mathbf{1}\big(|H_j|^{A/2}\le M_n^{A/2}\big) + \frac{H_j}{|H_j|}(M_n+R_j)^{A/2}\mathbf{1}\big(|H_j|^{A/2}>M_n^{A/2}\big)\,,$$

$j = 1, \dots, n$, $n = 1, 2, \dots$, but an inspection of the proof of Theorem 2.2 shows that the argument applies equally well to the latter triangular array, so that

$$\frac{1}{\big(nM_n^A P(|H_1| > M_n)\big)^{1/2}} \left(\sum_{j=1}^{n} X_{nj}^{\langle A/2 \rangle} - E\Big(\sum_{j=1}^{n} X_{nj}^{\langle A/2 \rangle} \Big) \right)$$

$$\Rightarrow \left(\frac{2A}{A-\alpha} \right)^{1/2} N(0,1)\,.$$

In particular, (extending the length of the rows of the triangular array) we see that

$$\frac{1}{\big(nM_n^A P(|H_1| > M_n)\big)^{1/2}} \left(\sum_{j=1}^{n} X_{nj}^{\langle A/2 \rangle} - \sum_{j=n+1}^{2n} X_{nj}^{\langle A/2 \rangle} \right)$$

$$\Rightarrow \left(\frac{4A}{A-\alpha} \right)^{1/2} N(0,1)\,.$$

Replacing n with $[n\gamma/2]$, we obtain (4.20) and, hence, finish the proof of part (i).

For part (ii), we define

$$b_n = \inf\big\{ x > 0 :\ P(|H_1|^{A/2} > x) \le n^{-1} \big\},\ n = 1, 2, \dots\,.$$

Then for some centering sequence (c_n) we have

$$b_n^{-1} \left(\sum_{j=1}^{n} H_j^{\langle A/2 \rangle} - c_n \right) \Rightarrow Y$$

with Y having a $S_{2\alpha/A}(\sigma, \beta, \mu)$ distribution with $\sigma^{2\alpha/A} = (C_{2\alpha/A})^{-1}$ and some β, μ; see Feller (1971). Because of the soft truncation, the triangular array $\big(X_{nj}^{\langle A/2 \rangle}\big)$ satisfies Theorem 2.1, and so

$$b_n^{-1} \left(\sum_{j=1}^{n} X_{nj}^{\langle A/2 \rangle} - c_n \right) \Rightarrow Y$$

with the same Y. Extending the rows of the triangular array gives us

$$b_n^{-1} \left(\sum_{j=1}^{n} X_{nj}^{\langle A/2 \rangle} - \sum_{j=n+1}^{2n} X_{nj}^{\langle A/2 \rangle} \right) \Rightarrow \left(\frac{2}{C_{2\alpha/A}} \right)^{A/(2\alpha)} S_1\,,$$

where S_1 is a symmetric $2\alpha/A$-stable random variable with unit scale. Replacing n with $[n\gamma/2]$ we obtain

$$\sum_{j=1}^{[\gamma n]} (-1)^j X_j^{\langle A/2 \rangle} \Rightarrow \left(\frac{\gamma}{C_{2\alpha/A}} \right)^{A/(2\alpha)} S_1\,. \tag{4.24}$$

Next, we also have

$$b_n^{-2}\sum_{j=1}^{n}|H_j|^A \Rightarrow \left(\frac{1}{C_{\alpha/A}}\right)^{A/\alpha} S_2\,,$$

where S_2 is a positive strictly α/A-stable random variable with unit scale; see once again Feller (1971). As before, because of the soft truncation, Theorem 2.1 applies, and we obtain

$$b_n^{-2}\sum_{j=1}^{n}|X_{nj}|^A \Rightarrow \left(\frac{1}{C_{\alpha/A}}\right)^{A/\alpha} S_2\,.$$

Replacing n with $(1-\gamma)n$, shows that

$$b_n^{-2}\sum_{j=[\gamma n]+1}^{n}|X_j|^A \Rightarrow \left(\frac{1-\gamma}{C_{\alpha/A}}\right)^{A/\alpha} S_2\,. \tag{4.25}$$

Since the numerator and the denominator of the statistic $Z_n(A;\gamma)$ in (4.17) are independent, the claim of part (ii) of the proposition follows from (4.24) and (4.25). □

Interestingly, the asymptotic distribution of the test statistic $Z_n(A;\gamma)$, under the null hypothesis, does not depend on the choice of the parameter A (as long as it an upper bound on the tail exponent α). Furthermore, under the null hypothesis this asymptotic distribution of the test statistic is light-tailed (e.g. some exponential moments are finite). On the other hand, the asymptotic distribution of the test statistic under the alternative is, clearly, heavy tailed, as even the second moment is infinite. Therefore, a reasonable test will reject the null hypothesis in favor of the alternative if the test statistic is too large. That is, we suggest the following test for the problem (4.16).

$$\text{reject } H_0 \text{ at significance level } p\in(0,1) \text{ if } \; Z_n(A;\gamma) > \frac{2\gamma}{1-\gamma}c_p\,, \tag{4.26}$$

with c_p such that $P(\chi_1^2 > c_p) = p$.

4.3. **Testing a stronger version of the hypothesis of hard truncation.** The test statistics $Z_n(A;\gamma)$ we used in the previous subsection for the problem (4.16) has a nondegenerate asymptotic distribution under both the null hypothesis and the alternative. This restricts the sensitivity of the resulting test. In order to obtain a more sensitive test we strengthen the null hypothesis. Specifically, in this subsection we consider the following problem of testing a null hypothesis against a simple alternative:

$$\left.\begin{array}{ll} H_0: & n^{1-\epsilon}P(|H_1|>M) \gg 1 \\ H_1: & nP(|H_1|>M) \ll 1 \end{array}\right\}, \tag{4.27}$$

where ϵ is a fixed number in $(0,1)$.

For this problem one can use the same test statistic $Z_n(A)$ defined in (4.3) as we used for the problem (4.1) of testing the hypothesis of soft truncation. Proposition 4.1 tells us that this test statistic diverges in probability to infinity under the hypothesis of hard truncation. The strengthened hypothesis of hard truncation in (4.27) allows us to quantify how fast this divergence takes place. This, in turn, can be used to build a test. The asymptotic distribution of $Z_n(A)$ under the hypothesis of soft truncation is described in Proposition 4.1. The next result provides an asymptotic distributional lower bound on the test statistic under the null hypothesis in the problem (4.27). As in the previous subsections, we assume that an upper bound (4.2) on the tail exponent is known.

Proposition 4.4. *Assume that $ER_1^{2A} < \infty$. Then under the strengthened hypothesis H_0 of hard truncation,*

$$\liminf_{n\to\infty} P\Big(n^{-\epsilon/2} Z_n(A) > x\Big) \ge e^{-x^2} \tag{4.28}$$

for every $x > 0$.

Proof. In the notation of the triangular array (1.2), consider the binomial random variable $N_n = \sum_{j=1}^n \mathbf{1}\big(|H_j| > M_n\big)$. The strengthened hypothesis of hard truncation implies that $P(N_n \ge n^\epsilon) \to 1$ as $n \to \infty$. Notice that, on an event of probability increasing to 1,

$$Z_n(A) \ge \frac{\sum_{j=1}^n (M_n + R_j)^A \mathbf{1}\big(|H_j| > M_n\big)}{\max_{j=1,\dots,n} (M_n + R_j)^A \mathbf{1}\big(|H_j| > M_n\big)}$$

$$\ge \frac{\sum_{j=1}^n R_j^A \mathbf{1}\big(|H_j| > M_n\big)}{\max_{j=1,\dots,n} R_j^A \mathbf{1}\big(|H_j| > M_n\big)} \,.$$

Therefore, for $x > 0$, using the assumption $ER_1^{2A} < \infty$, we have

$$\liminf_{n\to\infty} P\Big(n^{-\epsilon/2} Z_n(A) > x\Big) \ge \liminf_{n\to\infty} P\Big(\max_{j=1,\dots,N_n} R_j^A < n^{-\epsilon/2} N_n \frac{ER_1^A}{2} x^{-1}\Big)$$

$$\ge \liminf_{n\to\infty} E\left[\left(1 - \frac{x^2}{N_n}\right)^{N_n} \mathbf{1}\Big(N_n \ge n_\epsilon\Big)\right] \to e^{-x^2} \,,$$

as required. □

Proposition 4.4 tells us that under the hypothesis H_0, $n^{-\epsilon/2} Z_n(A)$ is, asymptotically, stochastically larger than the square root of the standard exponential random variable (independently of the parameter A). Therefore, we suggest the following test for the problem (4.27).

(4.29)

reject H_0 at significance level $p \in (0,1)$ if $Z_n(A) \le \big|\log(1-p)\big|^{1/2} n^{\epsilon/2}$.

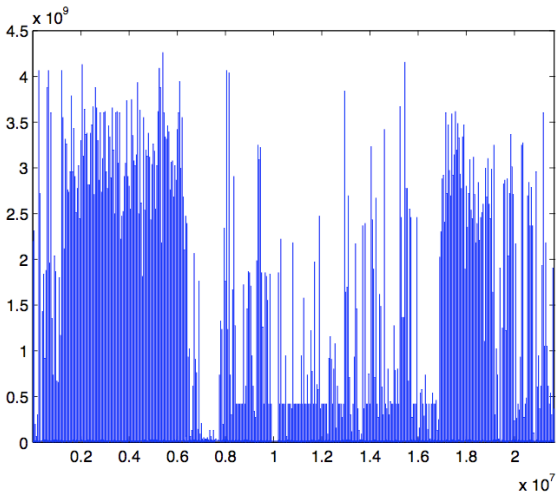

FIGURE 1. Think Times - the entire data set

5. Think Times and Object Sizes data: soft truncation or heavy truncation?

In this section we applied the statistical methods of Section 4 to two data sets. One data set contains "think times", or delays (in microseconds) between successive request/response exchanges between hosts using a TCP connection. The second data set contains the sizes (in bytes) of objects (files, HTTP responses, email messages, etc.) transferred on TCP connections. Both data sets were acquired by monitoring between 1:30 PM and 2:30 PM on July 24, 2006, the communication links connecting the site of a large commercial enterprize to the Internet. Both data sets exhibit visual evidence of heavy tails, and the Hill estimator confirms that (see below). Our goal is to check if the data sets show statistical evidence of soft or mild truncation of heavy tails.

5.1. **Think Times.** This data sets contains 2.1×10^7 observations which are plotted on Figure 1.

Clearly, the nature of this data set changes over time, and the nature of truncation of heavy tails may potentially change as well. In order to study this effect we have broken the data set into four pieces, with corresponding ranges $\left[0.11 \times 10^7, 0.64 \times 10^7\right]$; $\left[0.8 \times 10^7, 1.6 \times 10^7\right]$; $\left[1.7 \times 10^7, 1.9 \times 10^7\right]$ and $\left[1.95 \times 10^7, 2.1 \times 10^7\right]$. The individual pieces are plotted on Figure 2

The structure of the 4 individual pieces appears to be more stable than that of the entire data sets, and we proceed to analyze each piece separately. To do that, we first ran the Hill estimator with random k given in (3.4) on the first half of each of the 4 pieces. The estimation was conducted using $\beta, \gamma = 0.3, 0.4, 0.5, 0.6, 0.7$ and conservative upper bounds for α were obtained; these are presented in the following table.

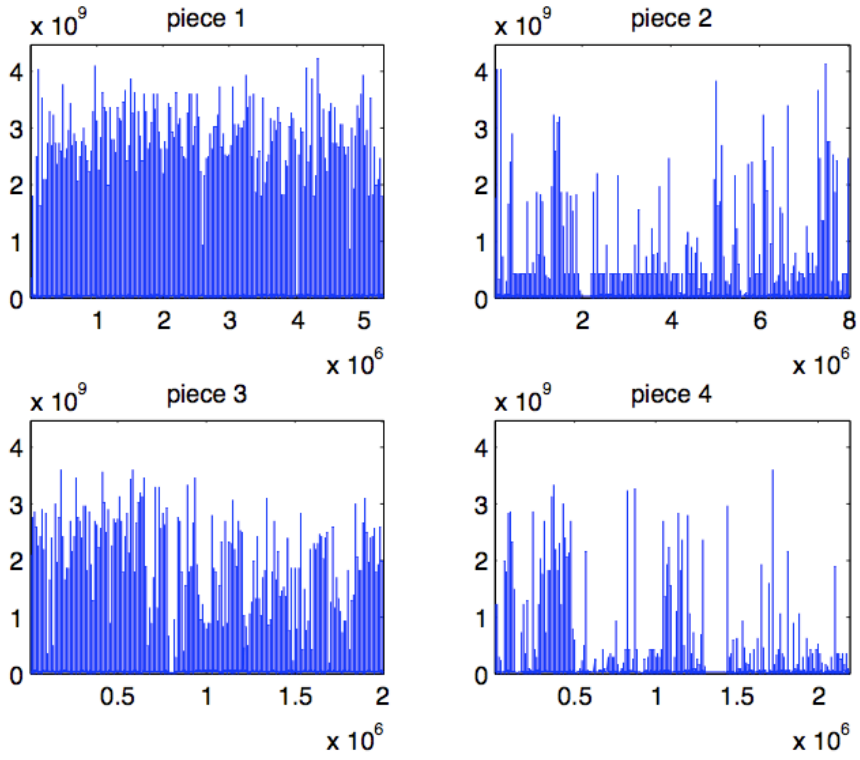

FIGURE 2. Think Times - the different pieces

piece	A
1	3.02
2	2.30
3	0.85
4	2.24

We then proceeded to use the second halves of each piece of the Think Times data set to test for soft and hard truncations.

Testing the hypothesis of soft truncation. The test statistic $Z_n(A_1)$ of Section 4.1 was computed for various values of A_1 larger than A. The results are reported in the following table.

A/A_1	piece 1	piece 2	piece 3	piece 4
0.5	31.43	5.81	154.05	3.57
0.6	51.59	7.99	205.37	4.72
0.7	77.39	10.74	271.27	6.11
0.8	108.08	14.20	361.74	7.81
0.9	142.78	18.57	491.31	9.91
0.95	161.38	21.16	576.73	11.13

Comparing the resulting values of the test statistic with the corresponding quantiles (or their upper bounds) of $Z(A/A_1)$, it is clear that the null hypothesis of soft truncation can be rejected for pieces 1 and 3. For piece 2, there is some evidence against the null hypothesis of hard truncation, while for piece 4 no such evidence exists.

Testing the hypothesis of hard truncation. The test statistic $Z_n(A;\gamma)$ of Section 4.2 was computed for various values of γ. The resulting p-values are reported in the following table.

γ	piece 1	piece 2	piece 3	piece 4
0.1	0.85	0.72	0.88	0.33
0.2	0.83	0.98	0.38	0.57
0.3	0.97	0.99	0.79	0.68
0.4	0.94	0.68	0.39	0.43
0.5	0.83	0.63	0.94	0.47
0.6	0.97	0.89	0.83	0.27
0.7	0.91	0.88	0.87	0.40
0.8	0.64	0.85	0.80	0.33
0.9	0.70	0.37	0.85	0.40

Clearly, the the hypothesis of hard truncation cannot be rejected for any of the four pieces.

Testing a stronger version of the hypothesis of hard truncation. The test statistics $Z_n(A)$ of Section 4.3 was computed and the corresponding p-values calculated for various values of ϵ. These are listed in the following table.

ϵ	piece 1	piece 2	piece 3	piece 4
0.1	1.00	1.00	1.00	1.00
0.2	1.00	1.00	1.00	1.00
0.3	1.00	1.00	1.00	0.91
0.4	1.00	0.73	1.00	0.45

It is clear that even the stronger version of the hypothesis of hard truncation cannot be rejected.

5.2. **Object Sizes.** This data set contains 2.2×10^7 observations. It is plotted in Figure 3. It does not appear that the nature of the observations changes with time, so we applied our statistical tests to the entire data set. After running the Hill estimator with random k and parameters β and γ as above, on the first half of the data set, we obtained a conservative upper bound on the value of the tail exponent α; this turned out to be $A = 1.69$. We used the second half of the Object Sizes data set to test for soft and hard truncations.

Testing the hypothesis of soft truncation. We evaluated the test statistic $Z_n(A_1)$ of Section 4.1 for a range of values of A_1 larger than A. The results are reported in the following table.

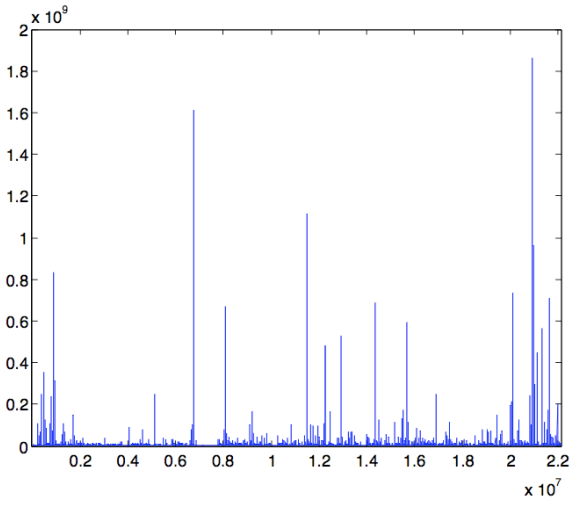

FIGURE 3. Data on Object Sizes

A/A_1	$Z_n(A_1)$
0.5	1.75
0.6	2.32
0.7	3.09
0.8	4.10
0.9	5.42
0.95	6.23

Comparing these with the corresponding quantiles (or their upper bounds) of $Z(A/A_1)$, we see that the hypothesis of soft truncation cannot be rejected.

Testing the hypothesis of hard truncation. We evaluated the test statistic $Z_n(A;\gamma)$ of Section 4.2 for various values of γ, and the obtained p-values are reported in the following table.

γ	p-value
0.1	0.50
0.2	0.36
0.3	0.73
0.4	0.77
0.5	0.95
0.6	0.94
0.7	0.94
0.8	0.97
0.9	0.72

The null hypothesis of hard truncation cannot be rejected.

Testing a stronger version of the hypothesis of hard truncation. We calculated the test statistics $Z_n(A)$ of Section 4.3 for various values of ϵ, and the p-values are given in the following table.

ϵ	p-value
0.1	1.00
0.2	0.86
0.3	0.33
0.4	0.08

The strengthened hypothesis of hard truncation becomes suspicious for $\epsilon = 0.4$, but overall our statistical tests do not produce clear evidence of the level of truncation for the Object Sizes data set.

6. Acknowledgment

The authors wish to thank Dr. F. Donelson Smith of the Network Research Laboratory in the Computer Science Department at the University of North Carolina at Chapel Hill for kindly providing the data sets analyzed in Section 5.

References

Aban, I., Meerschaert, M., and Panorska, A. (2006). Parameter estimation for the truncated pareto distribution. *Journal of the American Statistical Association*, 101(473):270–277.

Araujo, A. and Giné, E. (1980). *The Central Limit Theorem for Real and Banach Valued Random Variables*. Wiley, New York.

Asmussen, S. and Pihlsgard, M. (2005). Performance analysis with truncated heavy-tailed distributions. *Methodology and Computing in Applied Probability*, 7(4):439–457.

Barthelemy, P., Bertolotti, J., and Wiersma, S. (2008). A lévy flight for light. *Nature*, 453:495–498.

Bartumeus, F., da Luz, M., Vishwanathan, G., and Catalan, J. (2005). Animal search strategies: a quantitative random walk analysis. *Ecology*, 86(11):3078–3087.

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

Brockmann, D., Hufnagel, L., and Geisel, T. (2006). The scaling laws of human travel. *Nature*, 439:462–465.

Burrooughs, S. and Tebbens, S. (2001). Upper-truncated power laws in natural systems. *Pure and Applied Geophysics*, 158(4):741–757.

Corral, A. (2006). Universal earthquake-occurence jumps, correlations with time, and anomalous diffusion. *Physical Review Letters*, 97(17):178501.

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

Einmahl, J., Fils-Villetard, A., and Guillou, A. (2008). Statistics of extremes under random censoring. *Bernoulli*, 14(1):207–227.

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

Feller, W. (1971). *An Introduction to Probability Theory and its Applications*, volume 2. Wiley, New York, 2nd edition.

Gomez, C., Selman, B., Crato, N., and Kautz, H. (2000). Heavy–tailed phenomena in satisfiability and constraint satisfaction problems. *Journal of Automated Reasoning*, 24(1-2):67–100.

Gut, A. (2005). *Probability: A Graduate Course*. Springer.

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

Hong, S., Rhee, I., Kim, S., Lee, K., and Chong, S. (2008). Routing performance analysis of human-driven delay tolerant networks using the truncated levy walk model. In *Mobility Modles '08: Proceedings of the 1st ACM SIGMOBILE workshop on Mobility models*, pages 25–32, New York. ACM.

Hult, H., Lindskog, F., Mikosch, T., and Samorodnitsky, G. (2005). Functional large deviations for multivariate regularly varying random walks. *Annals of Applied Probability*, 15(4):2651–2680.

Jelenković, P. R. (1999). Network multiplexer with truncated heavy-tailed arrival streams. *INFOCOM '99. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, 2:625–632.

Maruyama, Y. and Murakami, J. (2003). Truncated Lévy walk of a nanocluster bound weakly to an atomically flat surface: Crossover from superdiffusion to normal diffusion. *Physical Review B*, 67(8):085406.

Microsoft Knowledge Base Article 154997 (2007). Description of the FAT32 file system.

Mikosch, T. (2009). *Non-Life Insurance Mathematics: An Introduction with the Poisson Process*. Springer, Berlin, 2nd edition.

Resnick, S. (1987). *Extreme Values, Regular Variation and Point Processes*. Springer-Verlag, New York.

Resnick, S. (2007). *Heavy-tail phenomena : probabilistic and statistical modeling*. Springer, New York.

Rosiński, J. (1990). On series representation of infinitely divisible random vectors. *The Annals of Probability*, 18:405–430.

Rvačeva, E. (1962). On domains of attraction of multi-dimensional distributions. *Selected Translations in Mathematical Statistics and Probability*, 2:183–205. Publisher: IMS-AMS.

Samorodnitsky, G. and Taqqu, M. (1994). *Stable Non-Gaussian Random Processes*. Chapman and Hall, New York.

Sato, K. (1999). *Lévy Processes and Infinitely Divisible Distributions*. Cambridge University Press.

Scholtz, C. and Contreras, J. (1998). Mechanics of continental rift architecture. *Geology*, 26(11):967–970.

Serrano, M., Flammini, A., and Menczer, F. (2009). Beyond zipf's law: Modeling the structure of human language. Technical Report.

Zaninetti, L. and Ferraro, M. (2008). On the truncated pareto distribution with applications. *Central European Journal of Physics*, 6(1):1–6.

School of Operations Research and Information Engineering, Cornell University, Ithaca, NY 14853, U.S.A.

E-mail address: ac427@cornell.edu

School of Operations Research and Information Engineering, Cornell University, Ithaca, NY 14853, U.S.A.

E-mail address: gennady@orie.cornell.edu