5-1-2010

Efficient Algorithms for Heavy-Tail Analysis under Interval Uncertainty

Vladik Kreinovich
University of Texas at El Paso, vladik@utep.edu

Monchaya Chiangpradit

Wararit Panichkitkosolkul

Comments:
Technical Report: UTEP-CS-10-09

Efficient Algorithms for Heavy-Tail Analysis under Interval Uncertainty

Vladik Kreinovich[1*],
Monchaya Chiangpradit[2],
and
Wararit Panichkitkosolkul[2]

[1]Department of Computer Science
University of Texas at El Paso
El Paso, TX 79968, USA
vladik@utep.edu

[2]Department of Applied Statistics
King Mongkut's University of Technology
North Bangkok, Bangkok 10800 Thailand
monchaya_c@hotmail.com, monchaya_c@yahoo.com
wararit_tu@hotmail.com, wararit@mathstat.sci.tu.ac.th

Abstract

Most applications of statistics to science and engineering are based on the assumption that the corresponding random variables are normally distributed, i.e., distributed according to Gaussian law in which the probability density function $\rho(x)$ exponentially decreases with x: $\rho(x) \sim \exp(-k \cdot x^2)$. Normal distributions indeed frequently occur in practice. However, there are also many practical situations, including situations from mathematical finance, in which we encounter heavy-tailed distributions, i.e., distributions in which $\rho(x)$ decreases as $\rho(x) \sim x^{-\alpha}$. To properly take this uncertainty into account when making decisions, it is necessary to estimate the parameters of such distributions based on the sample data $x_1, \ldots, x_n$ – and thus, to predict the size and the probabilities of large deviations. The most well-known statistical estimates for such distributions are the Hill estimator H for α and the Weismann estimator W for the corresponding quantiles.

These estimators are based on the simplifying assumption that the sample values x_i are known exactly. In practice, we often know the values

*Corresponding author.

x_i only approximately – e.g., we know the estimates $\widetilde{x}_i$ and we know the upper bounds Δ_i on the estimation errors. In this case, the only information that we have about the actual (unknown) value x_i is that x_i belongs to the interval $\mathbf{x}_i = [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$. Different combinations of values $x_i \in \mathbf{x}_i$ lead, in general, to different values of H and W. It is therefore desirable to find the ranges $[\underline{H}, \overline{H}]$ and $[\underline{W}, \overline{W}]$ of possible values of H and W. In this paper, we describe efficient algorithms for computing these ranges.

Keywords: heavy-tailed distributions, interval uncertainty, efficient algorithms, Hill estimator, Weissman estimator

1 Heavy-Tailed Distributions: Empirical Fact

Normal distributions: reminder. Most application of statistics to science and engineering are based on the assumption that the corresponding random variables are normally distributed, i.e., distributed according to Gaussian law in which the probability density function $\rho(x)$ exponentially decreases with x: $\rho(x) \sim \exp(-k \cdot x^2)$; see, e.g., [30].

Normal distributions indeed frequently occur in practice. This empirical fact can be justified by the Central Limit Theorem, according to which, under certain reasonable conditions, the joint effect of many relatively small factors is (approximately) normally distributed; see, e.g., [30].

Normal distribution in financial models. The quantitative study of stock prices can be traced back to a pioneering PhD dissertation of L. Bachelier [2] who has shown that for any fixed time quantum, the probabilities of stock price fluctuations of different size can be well described by a Gaussian random walk (what physicists call *Brownian motion*).

In the Gaussian random walk model, fluctuations of different sizes x are normally distributed. The Gaussian random walk model indeed describes small financial fluctuations reasonably well.

Limitations of the Gaussian description. While the Gaussian random walk model well describes the probabilities of *small* fluctuations, this model drastically underestimates the probabilities of *large* fluctuations. For example,

- in the normal distribution, fluctuations larger than 6σ have a negligible probability $\approx 10^{-8}$, while
- in real economic systems, even larger fluctuations occur every decade (and even more frequently).

It is important to properly take care of such fluctuations because when we underestimate the probability of large fluctuations, we thus underestimate risk – and become unprepared when large fluctuations occur.

Emergence of heavy-tailed (fractal) models. In the 1960s, Benoit Mandelbrot, the author of fractal theory, empirically studied the fluctuations and showed [17] that larger-scale fluctuations follow the power-law distribution, with the probability density function

$$\rho(x) = A \cdot x^{-\alpha}, \tag{1}$$

for some constant $\alpha \approx 2.7$.

In these distributions, the probability decreases much slower than for the normal distribution, so the tails are much heavier. Because of this, the asymptotically power-law distributions are also known as *heavy-tailed distributions*.

Heavy-tailed distributions are ubiquitous. The above empirical result, together with similar empirical discovery of power laws in other application areas, has led to the formulation of *fractal theory*; see, e.g., [18, 19].

Since then, similar asymptotically power-law distributions have been empirically found in other financial situations [4, 5, 7, 10, 20, 22, 29, 32, 33], and in many other application areas [3, 18, 21, 28].

2 Heavy-Tailed Distribution: Theoretical Justification

Need for a theoretical explanation. From the purely mathematical viewpoint, we can, in principle, have many different non-Gaussian distributions. In practice, however, asymptotically power law distributions are the most widely spread.

In many application areas, there is no good theoretical explanation for this empirical phenomenon. For example, in economics, these laws are not widely used by mainstream economists – exactly because they are empirical, they lack a clear economic justification [31]. Without a good theoretical explanation, economists are reluctant to rely on these laws being valid in the future as well – and to make serious decisions based on these laws.

As a result, in the existing financial decisions, economists often use Gaussian random walk models which are much less accurate than the empirically more accurate power law models. As a consequence of this practice, the existing financial instruments often underestimate the probability of large-scale crisis-type fluctuations.

To make financial systems more reliable and less vulnerable to large-scale crisis-style fluctuations, it is therefore important to overcome the economists' reluctance, and to provide a convincing theoretical explanation for the empirically observed power laws. This need is emphasized, e.g., in [31].

Most existing explanations are too complex. There exist several theoretical explanations for the empirical power law; see, e.g., [11]. These explanations are based on the deep mathematical analysis of complex systems. The complex

mathematical nature of these explanations makes them not very convincing for economists.

It is therefore desirable to provide simpler – and hopefully more convincing – explanations for the power law.

What we do in this section. In this section, we provide a detailed description of such a simpler explanation – based on scale invariance. The main ideas behind this explanation can be found in [24] and [16].

Analysis of the problem: a practice-oriented temporal reformulation of the probabilities. Ideally, we should be able to predict when the fluctuations will reach a given size x_0. In reality, as we have mentioned, economic fluctuations are random (unpredictable). As a result, we cannot predict the *exact* moment of time when fluctuations reach the threshold x_0. Instead, we can only predict the *average* time t before such a fluctuation occurs.

From this viewpoint, we would like to find the dependence $t(x_0)$ of this average time on the size of the fluctuation.

This dependence is naturally related to the probabilities. Indeed, the probability density function $\rho(x)$ means that the probability of a fluctuation of size x_0 is equal to $\rho(x_0)$. (To be more precise, it is equal to $\rho(x_0) \cdot h$, where h is the corresponding discretization step).

The probability that the fluctuation of this size occurs within a single time quantum Δt is equal to $\rho(x_0) \cdot h$. Thus, the expected number of such fluctuations during a single time quantum is $\rho(x_0) \cdot h$. During the time period t, we have $N \stackrel{\text{def}}{=} \dfrac{t}{\Delta t}$ time quanta. The expected number of fluctuations of size x_0 during this time period is therefore equal to

$$N \cdot (\rho(x_0) \cdot h) = \frac{t}{\Delta t} \cdot (\rho(x_0) \cdot h). \tag{2}$$

The average time $t(x_0)$ until such a fluctuation occurs can be estimated as the time t for which this expected number of fluctuations becomes close to 1:

$$\frac{t(x_0)}{\Delta t} \cdot (\rho(x_0) \cdot h) \approx 1, \tag{3}$$

hence

$$t(x_0) \approx \frac{\Delta t}{\rho(x_0) \cdot h}. \tag{4}$$

Thus, once we find the dependence $t(x)$, we will be able to find the desired probability density function $\rho(x)$ as

$$\rho(x) \approx \frac{\Delta t}{t(x) \cdot h} = \frac{\text{const}}{t(x)}, \tag{5}$$

where $\text{const} \stackrel{\text{def}}{=} \dfrac{\Delta t}{h}$.

Scale invariance: a natural requirement. We want to describe a general dependence $t(x)$ of the average time t during which the fluctuation of a given size occurs on the size x of this fluctuation.

When describing this dependence, one should take into account that the numerical value of the fluctuation size x depends on the choice of a measuring unit for describing fluctuations. In principle, different units can be chosen. For example, when the European countries changed from their original currencies to Euros, all the stock prices at local stock markets were accordingly re-scaled. In general, if instead of the original unit, we use a new unit which is λ times smaller, then the fluctuation whose size in the original unit is x has the value $x' = \lambda \cdot x$ in the new units.

It is reasonable to require that the expression describing dependence $t(x)$ should not depend on the choice of the unit. One needs to be careful, however, when formulating this natural requirement. Namely, we cannot simply assume that for the same numerical value x, the time is the same no matter which units we use. If we use a smaller unit than before, then

- a fluctuation whose size is one new unit is smaller than the fluctuation whose size is one original unit – and thus,
- the time to reach the 1 new unit size fluctuation should be smaller than the time to reach the 1 old unit size fluctuation.

So, to make a proper formalization, we must take into account that if we re-scale the units in which we measure fluctuations, we must accordingly change the units for time.

If we use the new unit for the fluctuation size, then instead of the numerical value x, we get a new numerical value $x' = \lambda \cdot x$. Thus, instead of the original time $t(x)$, we get a new time $t(x') = t(\lambda \cdot x)$. We require that this new time is actually the same time as $t(x)$, but expressed in different time units. If we denote the ratio of the corresponding time units by $r(\lambda)$, then we arrive at the formula

$$t(\lambda \cdot x) = r(\lambda) \cdot t(x). \tag{6}$$

Thus, we arrive at the following requirement: for every $\lambda > 0$, there exists a value $r(\lambda)$ for which, for all x and for all λ, we have

$$t(\lambda \cdot x) = r(\lambda) \cdot t(x). \tag{7}$$

Scale invariance implies power law. It is known that every continuous function $t(x)$ satisfying the above property has the form $t(x) = r \cdot x^\alpha$ for some α; see, e.g., [1], Section 3.1.1, or [24]. (This result was first proven in [26].)

Proof. For differentiable functions $t(x)$, the result about power functions is easy to prove. Indeed, if we differentiate both sides of (7) by λ and take $\lambda = 1$, we get

$$x \cdot \frac{dt}{dx} = \alpha \cdot t, \tag{8}$$

where $\alpha \stackrel{\text{def}}{=} r(1)$. By moving all the terms containing t into one side and all the terms containing x to the other side, we conclude that

$$\frac{dt}{t} = \alpha \cdot \frac{dx}{x}. \tag{9}$$

Integrating both sides, we get

$$\ln(t) = \alpha \cdot \ln(x) + c, \tag{10}$$

hence

$$t = e^{\alpha \cdot \ln(x) + c} = e^c \cdot \left(e^{\ln(x)}\right)^\alpha = C \cdot x^\alpha \tag{11}$$

for $C = e^c$.

Conclusion. Under a natural requirement that the distribution of economic fluctuations does not depend on the choice of a monetary unit, we conclude that $t(x) \sim x^\alpha$ and thus,

$$\rho(x) \sim \frac{\text{const}}{t(x)} \sim x^{-\alpha}. \tag{12}$$

Thus, the power law is justified.

3 Traditional Techniques of Heavy-Tail Analysis: Hill and Weissman Estimators

Need for heavy-tail analysis. As we have mentioned, there are many practical situations (including situations from mathematical finance), in which we encounter heavy-tailed distributions, i.e., distributions in which $\rho(x)$ decreases as $\rho(x) \sim x^{-\alpha}$.

To properly take this uncertainty into account when making decisions, it is necessary to estimate the parameters of such distributions based on the sample data $x_1, \ldots, x_n$ – and thus, to predict the size and the probabilities of large deviations.

Two main tasks of heavy-tail analysis. In view of the above, we have two tasks:

- First, based on the sample $x_1, \ldots, x_n$, we must be able to estimate the parameter α of the actual distribution.
- Second, we must estimate the size and probability of large deviations corresponding to this distribution.

Estimating the probability of large deviations: analysis of the problem. For every real number x_0, the probability $\text{Prob}(x > x_0)$ can be described as $1 - \text{Prob}(x \le x_0)$, i.e., as $1 - F(x_0)$, where $F(z) \stackrel{\text{def}}{=} \text{Prob}(x \le z)$ is the cumulative distribution function (CDF) of the corresponding probability distribution.

For the probability distribution withe the probability density $\rho(x) \sim x^{-\alpha}$, the corresponding CDF has the form

$$F(x) = \int^x \rho(z)\,dz \sim x^{-(\alpha-1)}. \tag{13}$$

An alternative way of describing the probabilities of large deviations comes from the fact that in practice, it is not possible to completely prevent the failure of a system. What we can do is decrease to the level that makes such a failure not realistically possible. For that purpose, in many practical applications, there is an *acceptable risk*, i.e., an allowable probability of failure p_0. Once this value is given, we may be interested to find the largest possible value x_0 within this risk, i.e., value x_0 for which the probability of exceeding this value is the "negligible" value p_0: $\text{Prob}(x > x_0) = p_0$. This condition can be equivalently described as $\text{Prob}(x \le x_0) = 1 - p_0$.

For each value $p \in [0, 1]$, the value x for which $F(x) = p$, is called the *p-th quantile* $Q(p)$ of the corresponding probability distribution. In these terms, the desired value x_0 is the $(1-p_0)$-th quantile: $x_0 = Q(1-p_0)$. Thus, in our second task, we need, given p_0, to find the estimate for the quantile $Q(1 - p_0)$.

Known solutions to the two tasks: Hill and Weissman estimators. Solutions to both estimations tasks are known; see, e.g., [28]:

- The most widely used solution to the first task is the Hill estimator first proposed by in [12]:

$$H = \frac{1}{k} \cdot \sum_{j=1}^{k} \ln(x_{(n-j+1)}) - \ln(x_{(n-k)}), \tag{14}$$

  where, as usual, $x_{(i)}$ denotes *order statistics*, i.e., the results of ordering the original sample in the increasing order:

$$x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}. \tag{15}$$

  This estimator estimates the value $\gamma \stackrel{\text{def}}{=} \dfrac{1}{\alpha - 1}$.

- The most widely used solution to the second task is the Weissman estimator for the quantile $Q(1 - p)$ first proposed in [34]:

$$W = x_{(n-k)} \cdot \left(\frac{k+1}{(n+1) \cdot p} \right)^H, \tag{16}$$

  where H is the Hill estimator.

Comment. To make our exposition clearer, let us briefly describe how both estimates can be derived.

Derivation of the Hill estimator: in brief. In general, in statistics, one of the most widely estimation method is the Maximum Likelihood method in which we select a distribution in which the probability (density) corresponding to the observed data is the largest possible; see, e.g., [30]. Moreover, it is known that under reasonable conditions, this method leads to asymptotically optimal estimates of the corresponding parameters.

Most widely used statistical estimates, such as the sample mean and sample variance, come from applying the Maximum Likelihood method to the corresponding normal distribution, with the unknown mean and variance σ^2.

In contrast to the normal distribution, when we know the expression of the probability density function (pdf) $\rho(x)$ for all x, in the heavy-tailed case, we only know the expression for the *asymptotic* values of $\rho(x)$, i.e., for the values $\rho(x)$ corresponding to the large x. Thus, instead of considering the whole sample, we only consider, for some reasonable value k, the $k+1$ largest values from this sample, i.e., values $x_{(n-k)}$, $x_{(n-k+1)}$, ..., $x_{(n)}$. (For recommendations on how to select k, the reader is referred, e.g., to [28]).

We therefore assume that on the semi-line $[x_{(n-k)}, \infty)$, the probability distribution has the form $\rho(x) \sim x^{-\alpha}$ for some constant α. Since we only know the distribution for $x \ge x_{(n-k)}$, we will therefore consider a *conditional* distribution – under the condition $x \ge x_{(n-k)}$. For this conditional distribution, the probability density $\rho_c(x)$ is equal to 0 for $x < x_{(n-k)}$ and it is equal to $\rho_c(x) = C \cdot x^{-\alpha}$ for $x \ge x_{(n-k)}$ for some constant C.

The value C can be determined from the condition that the overall probability is 1:

$$1 = \int_{x_{(n-k)}}^{\infty} \rho_c(x)\,dx = \int_{x_{(n-k)}}^{\infty} C \cdot x^{-\alpha}\,dx =$$

$$\frac{C}{-(\alpha-1)} \cdot x^{-(\alpha-1)}\Big|_{x_{(n-k)}}^{+\infty} = \frac{C}{\alpha-1} \cdot (x_{(n-k)})^{-(\alpha-1)}. \tag{17}$$

Thus,

$$C = (\alpha-1) \cdot (x_{(n-k)})^{\alpha-1}. \tag{18}$$

For each corresponding observation $x_{(n-k+1)}$, ..., $x_{(n-k+i)}$, ..., $x_{(n)}$ the corresponding probability density is $C \cdot (x_{(n-k+i)})^{-\alpha}$, so the overall probability $\mathcal{L}$ of observing x_i is equal to the product of these probabilities:

$$\mathcal{L} = C^k \cdot \prod_{i=1}^{k} (x_{(n-k+i)})^{-\alpha}. \tag{19}$$

Substituting the expression (18) into the formula (19), we conclude that

$$\mathcal{L} = (\alpha-1)^k \cdot \prod_{i=1}^{k} \left(\frac{x_{(n-k+i)}}{x_{(n-k)}}\right)^{-\alpha}. \tag{20}$$

According to the Maximum Likelihood method, we need to find the value α for which this value $\mathcal{L}$ is the largest possible. In the case of the Gaussian distribution, it is known that it is convenient to maximize the logarithm $\ln(\mathcal{L})$ instead of the original expression. (Maximizing logarithm is equivalent to maximizing the value since $\ln(z)$ is a monotonic function.)

Here too, using the logarithms simplifies the optimized expression into

$$\ln(\mathcal{L}) = k \cdot \ln(\alpha - 1) - \alpha \cdot \sum_{i=1}^{k} (\ln(x_{(n-k+i)}) - \ln(x_{(n-k)})). \tag{21}$$

Differentiating this expression with respect to α and equating the derivative to 0, we conclude that

$$\frac{k}{\alpha - 1} - \sum_{i=1}^{k} (\ln(x_{(n-k+i)}) - \ln(x_{(n-k)})) = 0, \tag{22}$$

i.e.,

$$\begin{aligned} \frac{1}{\alpha - 1} &= \frac{1}{k} \cdot \sum_{i=1}^{k} (\ln(x_{(n-k+i)}) - \ln(x_{(n-k)})) \\ &= \frac{1}{k} \cdot \sum_{j=1}^{k} \ln(x_{(n-j+1)}) - \ln(x_{(n-k)}). \end{aligned} \tag{23}$$

This is exactly the Hill estimator.

Derivation of the Weissman estimator: in brief. Based on the values $x_{(i)}$, we want to estimate the $(1-p)$-th quantile $Q(1-p)$.

In general, the i-th value $x_{(i)}$ out of n is an approximation for the quantile with $p_0 \approx \dfrac{i}{n}$. Thus, for a given p_0, we can estimate $Q(p_0)$ by the value $x_{(i)}$ for i for which $p_0 \approx \dfrac{i}{n}$, i.e., by the value $x_{(\lfloor p_0 \cdot n \rfloor)}$. In particular, for $p_0 = 1 - p$, we take $x_{(n-\lfloor p \cdot n \rfloor)}$, where $\lfloor z \rfloor$ (the "floor" function), as usual, means the largest integer which is smaller than or equal to z.

A more accurate estimation leads to $p_0 \approx \dfrac{i}{n+1}$ and $x_{(n-\lfloor p \cdot (n+1) \rfloor)}$.

This works well when $p \geq 1/n$, i.e., when we are estimating the probability of a deviation that has already been observed. However, the example of financial applications shows that it also desirable to predict the probability of large fluctuations that has not yet been observed. In other words, it is desirable to estimate the quantile $Q(1-p)$ for $p \ll 1/n$.

For the power law distribution with the probability density $\rho(x) \sim x^{-\alpha}$, we have $\text{Prob}(z > x) = C \cdot x^{-(\alpha-1)}$, i.e., using the Hill estimator notation $\gamma = 1/(\alpha - 1)$, $\text{Prob}(z > x) = C \cdot x^{-1/\gamma}$. Here:

- For the value $x_{(n-k)}$, as we have just mentioned, we have

$$\frac{k+1}{n+1} = C \cdot (x_{(n-k)})^{-1/\gamma}. \tag{24}$$

- For the desired quantile value $q \stackrel{\text{def}}{=} Q(1-p)$, we have

$$p = C \cdot q^{-1/\gamma}. \tag{25}$$

Dividing both sides of the first equation (24) by both sides of the second equation (25), we conclude that

$$\frac{k+1}{(n+1)\cdot p} = \left(\frac{q}{x_{(n-k)}}\right)^{1/\gamma}. \tag{26}$$

Raising both sides by the power γ, we get

$$\left(\frac{k+1}{(n+1)\cdot p}\right)^{\gamma} = \frac{q}{x_{(n-k)}}. \tag{27}$$

Now, multiplying both sides by $x_{(n-k)}$ and using the Hill estimator H for γ, we get

$$q = x_{(n-k)} \cdot \left(\frac{k+1}{(n+1)\cdot p}\right)^{H}, \tag{28}$$

which is exactly the Weissman estimator.

An alternative expression for the Weissman estimator. To compute a^b, a computer usually computes first the logarithm $b \cdot \ln(a)$ of this expression and then takes the exponent of this logarithm. Thus, from the computational viewpoint, it is therefore reasonable to get a simplified expression for $\ln(W)$.

By taking the logarithm of the expression (28), we conclude that

$$\ln(W) = \ln(x_{(n-k)}) + \ln\left(\frac{k+1}{(n+1)\cdot p}\right) \cdot H. \tag{29}$$

Substituting the expression (14) for the Hill estimator H into this formula, we conclude that

$$\ln(W) = \frac{1}{k} \cdot \ln\left(\frac{k+1}{(n+1)\cdot p}\right) \cdot \sum_{j=1}^{k} \ln(x_{(n-j+1)}) - \left(\ln\left(\frac{k+1}{(n+1)\cdot p}\right) - 1\right) \cdot \ln(x_{(n-k)}). \tag{30}$$

Monotonicity properties of the Hall estimator. From the original expression (14) for the Hill estimator, we can conclude that this expression has the following properties:

- the value H increases with $x_{(n-k+i)}$ for all $i = 1, \ldots, n$ (first monotonicity property);
- the value H decreases with $x_{(n-k)}$ (second monotonicity property); and
- under the condition that $x_{(n-k)}$ coincide with several consequent values

$$x_{(n-k)} = x_{(n-k+1)} = \ldots = x_{(n-k+s)}, \tag{31}$$

  the dependence of H on the common value $x_{(n-k)} = \ldots = x_{(n-k+s)}$ is also monotonic – i.e., increasing or decreasing (third monotonicity property).

The third monotonicity property follows from the fact that with respect to the logarithms $\ln(x_i)$, the Hill estimator H is a linear function, and a linear function is always either increasing or decreasing with respect to each of its variables – depending on whether a coefficient at this variable is non-negative of non-positive. Since $\ln(z)$ is an increasing function, monotonicity with respect to $\ln(x_i)$ implies monotonicity with respect to x_i as well.

Monotonicity properties of the Weissman estimators. Based on the expression (30), we conclude that the Weissman estimator has the same monotonicity properties. Indeed, since $p \ll 1/n$, we have $\dfrac{k+1}{(n+1)\cdot p} \gg 1$ and thus, $\ln\left(\dfrac{k+1}{(n+1)\cdot p}\right) > 0$ and $\ln\left(\dfrac{k+1}{(n+1)\cdot p}\right) - 1 > 0$. Thus, the Weissman estimator W satisfies the same three properties:

- the value W strictly increases with $x_{(n-k+i)}$ for all $i = 1, \ldots, n$ (first monotonicity property);
- the value W strictly decreases with $x_{(n-k)}$ (second monotonicity property); and
- under the condition that $x_{(n-k)}$ coincide with several consequent values

$$x_{(n-k)} = x_{(n-k+1)} = \ldots = x_{(n-k+s)}, \tag{32}$$

  the dependence of W on the common value $x_{(n-k)} = \ldots = x_{(n-k+s)}$ is also monotonic – i.e., increasing or decreasing (third monotonicity property).

Comment. In the following text, we will consider arbitrary estimators that satisfy these three monotonicity properties. Thus, our results apply not only to the Hill and Weissman estimators, but also to all other estimators that satisfy these properties.

4 Need to Take Interval Uncertainty into Account

Existing estimators: implicit assumption. The above estimators are based on the simplifying assumption that the sample values x_i are known exactly.

Possibility of interval uncertainty. In practice, we often know the values x_i only approximately. In other words, instead of the *exact* value of x_i, we only know the *approximate* estimation $\widetilde{x}_i$. We also have some information about the approximation error $\Delta x_i \stackrel{\text{def}}{=} \widetilde{x}_i - x_i$.

In some cases, we know the probability distribution of different values of the approximation error. However, in many practical situations, we only the upper bound Δ_i on this error, i.e., the value for which $|\Delta x_i| \le \Delta_i$; see, e.g., [27].

In such situations, the only information that we have about the actual (unknown) value x_i is that x_i belongs to the interval

$$\mathbf{x}_i = [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i].$$

Because of this, such uncertainty is also known as an *interval uncertainty.*

Interval uncertainty in financial problems. Interval uncertainty also naturally appears in the analysis of financial data; see, e.g., [13] and references therein. For example, in the analysis of stock market data, each sample value x_i may represent the price of a certain stock on the i-th day. In reality, the price of each stock slightly fluctuates during the day.

Usually, practitioners take, as x_i, the average price or the price at a certain specific time. The problem is that there are several possibilities of select a single day price, and different selections lead to (slightly) different results. It is therefore reasonable, instead of artificially picking one number x_i, to consider the entire interval $[\underline{x}_i, \overline{x}_i]$ of all possible prices offered during the i-th day.

As shown in [13], not only this approach more reasonable – the resulting use of the additional information about daily variances of stock prices leads to a better predictions of future stock values.

Need to take interval uncertainty into account: formulation of the problem. Different combinations of values $x_i \in \mathbf{x}_i$ lead, in general, to different values of the estimators H and W.

It is therefore desirable to find the ranges $[\underline{H}, \overline{H}]$ and $[\underline{W}, \overline{W}]$ of possible values of H and W:

$$\mathbf{H} = [\underline{H}, \overline{H}] = \{H(x_1, \ldots, x_n) : x_1 \in \mathbf{x}_1, \ldots, x_n \in \mathbf{x}_n\}; \tag{33}$$

$$\mathbf{W} = [\underline{W}, \overline{W}] = \{W(x_1, \ldots, x_n) : x_1 \in \mathbf{x}_1, \ldots, x_n \in \mathbf{x}_n\}. \tag{34}$$

This problem is a particular case of interval computations. Due to the ubiquity of interval uncertainty, the need to estimate a range of a given function $f(x_1, \ldots, x_n)$ over given intervals $\mathbf{x}_1, \ldots, \mathbf{x}_n$ occurs in many other application areas. The problem of computing this range is known as the main problem of *interval computations*; see, e.g., [14, 23].

Interval computations is, in general, NP-hard. In spite of the simplicity of the problem's formulation, in general, the interval computations problem is NP-hard (computationally intensive [25]); see, e.g., [15].

It is even NP-hard if we restrict ourselves to simple functions: e.g., to quadratic ones. Moreover, the problem is NP-hard even for the simplest statistically meaningful quadratic function: the function $V(x_1, \ldots, x_n)$ that describes the sample variance [8, 9]

$$V(x_1, \ldots, x_n) = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n} \cdot \sum_{i=1}^{n} x_i \right)^2. \tag{35}$$

One may think that since, e.g., the Hill estimator is easier than a quadratic function – once we get to logarithms, it only uses addition and sorting (i.e., min and max operations) – however, it is known that even if only limit ourselves to min and max, the problem of interval computations still remains NP-hard.

What we do in this paper. In this paper, we show that for the Hill and Weissman estimators, there are efficient algorithms for computing their ranges over given intervals.

In other words, in this paper, we produce efficient algorithms for heavy-tail analysis under interval uncertainty.

5 Computing Ranges of Monotonic Estimators under Interval Uncertainty: Analysis of the Problem

What we do in this section. In order to develop efficient algorithms for estimating the ranges of the Hill and Weissman estimators, let us analyze the problem of computing this range. To perform this analysis, let us first recall the properties of these estimators and what exactly we want to compute.

General description of the problem. To describe a general monotonic estimator, we first need to define an auxiliary function $E(z_0, z_1, \ldots, z_k)$.

Let k be a positive integer, and let $E(z_0, z_1, \ldots, z_k)$ be a function which is defined for all the tuples that satisfy the inequalities

$$z_0 \le z_1 \le \ldots \le z_k \tag{36}$$

and which satisfies the following three properties:

- the function E is increasing in each of the variables $z_1, \ldots, z_k$;
- the function E is decreasing as a function of z_0; and
- for every s, the function

$$E_s(z_0, z_{s+1}, \ldots, z_k) \stackrel{\text{def}}{=} E(z_0, z_0, \ldots, z_0, z_{s+1}, \ldots, z_k) \tag{37}$$

  is either increasing or decreasing.

In terms of the auxiliary function $E(z_0, z_1, \ldots, z_k)$, the estimator $F(x_1, \ldots, x_n)$ can be defined as follows: For every sample $x_1, \ldots, x_n$, we find $k+1$ largest values $x_{(n-k)}, \ldots, x_{(n)}$, and compute the value

$$F(x_1, \ldots, x_n) \stackrel{\text{def}}{=} E(x_{(n-k)}, x_{(n+1-k)}, \ldots, x_{(n)}). \tag{38}$$

Now, instead of the exact values $x_1, \ldots, x_n$, we only know interval $\mathbf{x}_i = [\underline{x}_i, \overline{x}_i]$ of possible values. For each combination of values $x_i \in \mathbf{x}_i$, we can apply the formula (38) and produce the corresponding estimate $F(x_1, \ldots, x_n)$. Our objective is to find the range

$$\mathbf{F} = [\underline{F}, \overline{F}] = \{F(x_1, \ldots, x_n) : x_i \in \mathbf{x}_i\} \tag{39}$$

of possible values of F when $x_i \in \mathbf{x}_i$.

A comment about notations. In accordance with the usual notations of order statistics, by (i), for each selection $x_i \in \mathbf{x}_i$, we will denote the index of the value which is i-th in the increasing order in the ordering of the selected values

$$x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}. \tag{40}$$

Thus, $\underline{x}_{(i)}$ means the $\underline{x}_j$ for $j = (i)$.

Warning. The order of the lower endpoints $\underline{x}_i$ and the order of the upper endpoints $\overline{x}_i$ do not necessarily follow the order of the values themselves. Thus, while we always have, e.g., $x_{(1)} \le x_{(2)}$, we may have $\underline{x}_{(1)} > \underline{x}_{(2)}$. This happens because the notation (i) reflects the order of the values x_i but *not* of the lower endpoints. Thus, $\underline{x}_{(1)}$ does not mean the smallest of the lower endpoints – it simply means the lower endpoint corresponding to the smallest value.

An auxiliary result about the optimizing values of $x_{(n-k)}$. In our formulation, we have n intervals $[\underline{x}_i, \overline{x}_i]$ and thus, $2n$ endpoints $\underline{x}_i$ and $\overline{x}_i$.

Let us prove the following two properties:

- the first property is that for every tuple $x = (x_1, \ldots, x_n)$, there exists another tuple $x' = (x'_1, \ldots, x'_n)$ with $x'_i \in \mathbf{x}_i$ in which
  - the value $x'_{(n-k)}$ is equal to one of the endpoints, and

- we have $F(x'_1, \ldots, x'_n) \geq F(x_1, \ldots, x_n)$;

- the second property is that for every tuple $(x_1, \ldots, x_n)$, there exists another tuple $(x'_1, \ldots, x'_n)$ with $x'_i \in \mathbf{x}_i$ in which
  - the value $x'_{(n-k)}$ is equal to one of the endpoints, and
  - we have $F(x'_1, \ldots, x'_n) \leq F(x_1, \ldots, x_n)$.

As a result of these two properties, to find both

- the smallest possible value $\underline{F}$ of the function $F(x_1, \ldots, x_n)$ and
- the largest possible value $\overline{F}$ of this function,

it is sufficient to consider tuples for which $x_{(n-k)}$ is equal to one of the endpoints.

Proving the first property. Let us first prove the first property – the one that enables us to make a conclusion about the maximum.

Let us start with an arbitrary tuple x. For this tuple, the value $x_{(n-k)}$ belongs to the corresponding interval $[\underline{x}_{(n-k)}, \overline{x}_{(n-k)}]$ and is, therefore, larger than or equal to the lower endpoint of this interval $\underline{x}_{(n-k)} \leq x_{(n-k)}$.

Thus, there exists an endpoint which is smaller than or equal to $x_{(n-k)}$. Let z denote the largest of all the endpoints which are smaller than or equal to $x_{(n-k)}$. Since all these endpoints are smaller than or equal to $x_{(n-k)}$, the largest of them is also smaller than or equal to $x_{(n-k)}$: $z \leq x_{(n-k)}$.

Let us define the new tuple x'_i as follows:

- for all $j = 1, \ldots, k$, we take $x'_{(n-k+j)} = x_{(n-k+j)}$; in other words, we keep all the value $x_{(n-k+1)}$, $\ldots$, $x_{(n-k)}$ intact;
- for all $j \leq n-k$, we take $x'_{(j)} = \min(z, x_{(j)})$.

In particular, since $z \leq x_{(n-k)}$, the above selection means that we take $x'_{(n-k)} = z$.

Let us show that this new tuple x' is the desired one.

Indeed, by our choice, the new value $x'_{(n-k)}$ coincides with one of the endpoints z.

Now, in the new tuple, $x'_{(n-k)} = z \leq x_{(n-k)}$, and all the values $x_{(n-k+1)}, \ldots, x_{(n-k)}$ remain intact. Since the function F is a decreasing function of $x_{(n-k)}$, we can thus conclude that $F(x'_1, \ldots, x'_n) \geq F(x_1, \ldots, x_n)$.

So, to complete our proof, it is sufficient to show that the new values $x'_{(j)}$, $j \leq n-k$, still belong to the corresponding intervals $[\underline{x}_{(j)}, \overline{x}_{(j)}]$.

By definition of the minimum, each new value $x'_{(j)} = \min(z, x_{(j)})$

- is either equal to the old value $x_{(j)}$), or
- it is equal to z.

If the new value $x'_{(j)}$ is equal to the old value $x_{(j)}$, then, of course, it belongs to the same interval $[\underline{x}_{(j)}, \overline{x}_{(j)}]$ as the old value. Let us now consider the case when the minimum is equal to z, i.e., when $x'_{(j)} = z < x_{(j)}$. We need to prove that the new value $x'_{(j)}$ belongs to the interval $[\underline{x}_{(j)}, \overline{x}_{(j)}]$, i.e.:

- that $x'_{(j)} \le \overline{x}_{(j)}$, and
- that $\underline{x}_{(j)} \le x'_{(j)}$.

The first inequality is the easiest to prove: we know that $x_{(j)} \le \overline{x}_{(j)}$, so from $x'_{(j)} < x_{(j)}$, we conclude that indeed $x'_{(j)} \le \overline{x}_{(j)}$.

To prove the second inequality, let us recall that the original value $x_{(j)}$ belongs to the corresponding interval $[\underline{x}_{(j)}, \overline{x}_{(j)}]$ and therefore, $\underline{x}_{(j)} \le x_{(j)}$. Since $j \le n-k$ and the values $x_{(j)}$ are ordered by the index, we thus conclude that $x_{(j)} \le x_{(n-k)}$. By transitivity of order, from $\underline{x}_{(j)} \le x_{(j)}$ and $x_{(j)} \le x_{(n-k)}$, we conclude that $\underline{x}_{(j)} \le x_{(n-k)}$. Thus, $\underline{x}_{(j)}$ is one the endpoints which is smaller than or equal to $x_{(n-k)}$. Since z is the largest of such endpoints, we thus have $\underline{x}_{(j)} \le z$, i.e., $\underline{x}_{(j)} \le x'_{(j)}$.

So, each of the new values $x'_{(j)}$ does belong to the corresponding interval $[\underline{x}_{(j)}, \overline{x}_{(j)}]$.

The first property is proven.

Proving the second property. Let us now prove the second property. Let x be an arbitrary tuple. Due to sorting, the value $x_{(n-k)}$ is smaller than or equal to all the following values:

$$x_{(n-k)} \le x_{(n-k+1)} \le \ldots \le x_{(n)}. \tag{41}$$

It may happen that in the tuple x, the value $x_{(n-k)}$ is equal to some of the following values; let us denote by s the number of such terms. Then,

$$x_{(n-k)} = x_{(n-k+1)} = \ldots = x_{(n-k+s)} < x_{(n-k+s+1)} \le \ldots \le x_{(n-k)}. \tag{42}$$

If $x_{(n-k)}$ is equal to one of the endpoints, then we can simply take $x'_i = x_i$.

If $x_{(n-k)}$ is not equal to one of the endpoints, then we will select x' by changing all the values $x_{(n-k)}$, $x_{(n-k+1)}$, $\ldots$, $x_{(n-k+s)}$ at the same time, i.e., by taking

$$x'_{(n-k)} = x'_{(n-k+1)} = \ldots = x'_{(n-k+s)} = z \tag{43}$$

for some value z.

Because of the third monotonicity property of the estimator $F(x_1, \ldots, x_n)$, if we only change these $s+1$ equal values without changing the larger values $x_{(n-k+s+1)}$, $\ldots$, $x_{(n-k)}$, the function F becomes either an decreasing or an increasing function of z. Let us consider these two cases one by one.

If the function F is decreasing, then, similarly to the proof of the first property, we can decrease all these common values, including $x_{(n-k)}$, to the largest possible endpoint $z \le x_{(n-k)}$, and thus, get a new tuple x' in which $x'_{(n-k)}$ is equal to one of the endpoints and $F(x') \le F(x)$.

If the function F is increasing, we will increase z as much as possible without violating the order and the conditions $x_{(j)} \in [\underline{x}_{(j)}, \overline{x}_{(j)}]$. When we increase z, the value of the function F increases or remains the same.

To avoid violating the order, we must have $z \le x_{(n-k+s+1)}$. To avoid getting out of the corresponding intervals, we must have $z \le \overline{x}_{(n-k)}$, $z \le \overline{x}_{(n-k+1)}$, $\ldots$, $z \le \overline{x}_{(n-k+s)}$. The largest possible value z that bounded by all these bounds is the smallest of these bounds. Thus, we will take

$$z = \min(x_{(n-k+s+1)}, \overline{x}_{(n-k+1)}, \ldots, \overline{x}_{(n-k+s)}). \tag{44}$$

If the minimum (44) is equal to one of the endpoints, then we get the desired tuple x', with $x'_{(n-k)} = z$ equal to one of the endpoints, and with $F(x') \ge F(x)$.

If the minimum (44) is equal to $x_{(n-k+s+1)}$, then we get ourselves a new tuple in which we have $s+1$ values equal to $x_{(n-k)}$. We then repeat the process for the new tuple, etc.

At each step,

- either we reach a tuple for which $x'_{(n-k)}$ is equal to an endpoint,
- or we increase the number of equal values after $x_{(n-k)}$.

Since there are only k values after $x_{(n-k)}$, and on each iteration, the number of equal values increases by 1, after $\le k+1$ iterations, we reach a tuple x' in which $x'_{(n-k)}$ is equal to an endpoint.

At each step, we decrease the value of F or keep it intact. Thus, the second property is proven too.

Main idea of our algorithm. Due to our auxiliary statement, to find the maximal or minimal value of $f(x_1, \ldots, x_n)$, it is sufficient, for each of $2n$ endpoints t, to consider all the tuples for which $x_{(n-k)} = t$.

For each of the endpoints t, we then find the smallest $\underline{F}(t)$ and the largest $\overline{F}(t)$ of the values $F(x_1, \ldots, x_n)$ under the condition that $x_{(n-k)} = t$.

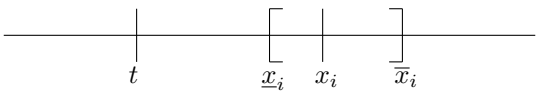
Once we find these values, we then take:

- the largest of the values $\overline{F}(t)$ as $\overline{F}$, and
- the smallest of the values $\underline{F}(t)$ as $\underline{F}$.

The remaining question is: how do we compute the values $\underline{F}(t)$ and $\overline{F}(t)$?

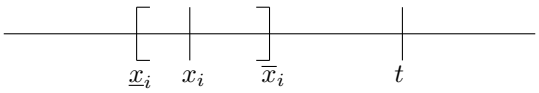
How to compute $\underline{F}(t)$ and $\overline{F}(t)$: preliminary analysis. For each endpoint $t = x_{(n-k)}$ and for each interval $[\underline{x}_i, \overline{x}_i]$, we have three possible situations:

- If $t < \underline{x}_i$, this means that every value $x_i \in [\underline{x}_i, \overline{x}_i]$ is also larger than t.



Thus, no matter which value x_i we choose from the corresponding interval $[\underline{x}_i, \overline{x}_i]$, this value x_i will be larger than $t = x_{(n-k)}$. So, we will have $x_i = x_{(n-k+j)}$ for some $j > 0$.

- If $t > \overline{x}_i$, this means that every value $x_i \in [\underline{x}_i, \overline{x}_i]$ is also smaller than t.



Thus, no matter which value x_i we choose from the corresponding interval $[\underline{x}_i, \overline{x}_i]$, this value x_i will be smaller than $t = x_{(n-k)}$. So, we will have $x_i = x_{(j)}$ for some $j < n - k$.

- Finally, if $\underline{x}_i \le t \le \overline{x}_i$, i.e., if t belongs to the corresponding interval $[\underline{x}_i, \overline{x}_i]$, then a value x_i selected from this interval can be both below and above $x_{(n-k)}$ in the sorting $x_{(j)}$.

So, once we selected the endpoint t as the value $x_{(n-k)}$, let us first count

- how many values x_i are guaranteed to be above this value $t = x_{(n-k)}$ and
- how many values x_i are guaranteed to be below this value $t = x_{(n-k)}$.

Let t^+ denote the number of all the indices i for which $t < \underline{x}_i$, and let t^- denote the number of all the indices i for which $t > \overline{x}_i$.

Overall, we have no more than k values $x_{(j)}$ which are larger then $x_{(n-k)}$, and no more than $n - k - 1$ values $x_{(j)}$ which are smaller then $x_{(n-k)}$. Thus, if $t^+ > k$ or $t^- > n - k - 1$, the choice of t as $x_{(n-k)}$ is simply impossible, so this endpoint t should be simply dismissed from our computations.

Let us now concentrate on the remaining endpoints, for which $t^+ \le k$ and $t^- \le n - k - 1$. Due to the the fact that F is an increasing function of each of the variables $x_{(n-k+j)}$, for t^+ indices x_i for which $t < \underline{x}_i$, we select:

- the largest possible value $\overline{x}_i$ when we compute $\overline{F}(t)$, and
- the smallest possible value $\underline{x}_i$ when we compute $\underline{F}(t)$.

For the remaining $k - t^+$ values $x_{(n-k+j)}$, we must select some of the $n - t^+ - t^-$ indices x_i for which $t \in [\underline{x}_i, \overline{x}_i]$.

To find the maximum $\overline{F}(t)$, for each of the selected indices i, we should similarly take $x_i = \overline{x}_i$. Thus, we should select $k - t^+$ largest of these values $\overline{x}_i$. (To find these largest values for all t, it makes sense to pre-sort the upper endpoints $\overline{x}_i$.)

To find the minimum $\underline{F}(t)$, we should select the smallest values which are still $\ge t$ – i.e., the values equal to t.

Thus, we arrive at the following algorithm for computing $\underline{F}$ and $\overline{F}$.

6 Algorithm for Computing Ranges of Monotonic Estimators under Interval Uncertainty

Formulation of the problem: brief reminder. We need to compute the endpoints $\underline{F}$ and $\overline{F}$ of the range (39) of a given monotonic estimator $F(x_1, \ldots, x_n)$ (formula (38)) when each variable x_i belongs to the known interval $[\underline{x}_i, \overline{x}_i]$.

Preliminary step. We sort all the upper endpoints $\overline{x}_i$ into an increasing sequence.

Computing $\overline{F}$. To compute $\overline{F}$, we repeat the following procedure for each of $2n$ endpoints $t = \underline{x}_i$ and $t = \overline{x}_i$, $i = 1, \ldots, n$.

- First, by considering indices $i = 1, \ldots, n$, we count
  - the number t^+ of all the indices i for which $t < \underline{x}_i$, and
  - the number t^- of all the indices i for which $t > \overline{x}_i$.
- If $t^+ > k$ or $t^- > n - k - 1$, we dismiss the endpoint t and move to a next one. Otherwise:
  - for all t^+ indices i for which $t < \underline{x}_i$, we select $x_i = \overline{x}_i$;
  - from the $n - t^+ - t^-$ indices for which $\underline{x}_i \le t \le \overline{x}_i$, we select $k - t^+$ indices with the largest values of $\overline{x}_i$; for these values, we take $x_i = \overline{x}_i$;
  - for one of the remaining indices, we take $x_i = t$;
  - for all other remaining indices, we take $x_i = \underline{x}_i$.
- Then, we compute the value $F(x_1, \ldots, x_n)$ for the selected x_i; we denote this value as $\overline{F}(t)$.

Finally, we compute the largest of the $\le 2n$ values $\overline{F}(t)$ as the desired value $\overline{F}$.

Computing $\underline{F}$. To compute $\underline{F}$, we repeat the following procedure for each of $2n$ endpoints $t = \underline{x}_i$ and $t = \overline{x}_i$, $i = 1, \ldots, n$.

- First, by considering indices $i = 1, \ldots, n$, we count
  - the number t^+ of all the indices i for which $t < \underline{x}_i$, and
  - the number t^- of all the indices i for which $t > \overline{x}_i$.
- If $t^+ > k$ or $t^- > n - k - 1$, we dismiss the endpoint t and move to a next one. Otherwise:
  - for all t^+ indices i for which $t < \underline{x}_i$, we select $x_i = \underline{x}_i$;
  - from the $n - t^+ - t^-$ indices for which $\underline{x}_i \le t \le \overline{x}_i$, we select $k + 1 - t^+$ indices for which we take $x_i = t$;

– for the remaining indices, we take $x_i = \underline{x}_i$.

- Then, we compute the value $F(x_1, \ldots, x_n)$ for the selected x_i; we denote this value as $\underline{F}(t)$.

Finally, we compute the smallest of the $\le 2n$ values $\underline{F}(t)$ as the desired value $\underline{F}$.

Computational complexity of this algorithm: analysis. The first step – sorting – requires $O(n \cdot \log_2(n))$ steps; see, e.g., [6].

In the main part of the algorithm, for each of $2n$ endpoints t, we process each of the n intervals $[\underline{x}_i, \overline{x}_i]$ by using a constant number of computational steps, and then compute F once. Thus, for each endpoint t, we perform $O(n)$ computations and one call to f.

Overall, in the main part, we thus need $n \cdot O(n) = O(n^2)$ computations and $O(n)$ calls to F. Thus, the total computation time of our algorithm is

$$O(n \cdot \log_2(n)) + O(n^2) + O(n) \cdot t_F, \tag{45}$$

where t_F is a time needed for a single call to the function $F(x_1, \ldots, x_n)$.

Since $n \cdot \log_2(n) \le n^2$, we have $O(n \cdot \log_2(n)) + O(n^2) = O(n^2)$. Hence, we arrive at the following conclusion:

Computational complexity of this algorithm: result. The above algorithm requires time

$$O(n^2) + O(n) \cdot t_F, \tag{46}$$

where t_F is a time needed for a single call to the function $F(x_1, \ldots, x_n)$.

7 Algorithm: Numerical Example

Interval data. Let us assume that we have $n = 5$ interval data points:

$$[\underline{x}_1, \overline{x}_1] = [e^0, e^3]; \quad [\underline{x}_2, \overline{x}_2] = [e^1, e^2]; \quad [\underline{x}_3, \overline{x}_3] = [e^1, e^3];$$
$$[\underline{x}_4, \overline{x}_4] = [e^2, e^3]; \quad [\underline{x}_5, \overline{x}_5] = [e^2, e^4], \tag{47}$$

and we want to find the range of the Hill estimator corresponding to $k = 2$.

Comment. We have selected all the endpoints $\underline{x}_i$ and $\overline{x}_i$ to be of the type e^z with a simple z, so that their logarithms – which are used in the expression for the Hill estimator – becomes easy to compute and process.

Preliminary step. According to our algorithm, at this step, we sort the intervals in the increasing order of their upper endpoints $\overline{x}_i$. As a result, we get the following reordering of the intervals:

$$[\underline{x}_1, \overline{x}_1] = [e^1, e^2]; \quad [\underline{x}_2, \overline{x}_2] = [e^0, e^3]; \quad [\underline{x}_3, \overline{x}_3] = [e^1, e^3];$$
$$[\underline{x}_4, \overline{x}_4] = [e^2, e^3]; \quad [\underline{x}_5, \overline{x}_5] = [e^2, e^4]. \tag{48}$$

Computing $\overline{H}$. According to our algorithm, we consider all endpoints t. In our case, there are five different endpoints: $t = e^0$, $t = e^1$, $t = e^2$, $t = e^3$, and $t = e^4$. Let us consider these endpoints one by one.

Computing $\overline{H}$: case of $t = e^0$. First, we consider the value $t = e^0$. In this case, we have $t^+ = 4$ intervals for which $t < \underline{x}_i$: intervals

$$[\underline{x}_1, \overline{x}_1] = [e^1, e^2], \quad [\underline{x}_3, \overline{x}_3] = [e^1, e^3], \quad [\underline{x}_4, \overline{x}_4] = [e^2, e^3],$$

$$[\underline{x}_5, \overline{x}_5] = [e^2, e^4].$$

Since $4 = t^+ > k = 2$, we dismiss this endpoint.

Computing $\overline{H}$: case of $t = e^1$. Then, we consider the value $t = e^1$. In this case:

- We have $t^+ = 2$ intervals for which $t < \underline{x}_i$: intervals

  $$[\underline{x}_4, \overline{x}_4] = [e^2, e^3] \text{ and } [\underline{x}_5, \overline{x}_5] = [e^2, e^4];$$

  here, $t^+ \le k$.
- We also have $t^- = 0$ intervals for which $\overline{x}_i < t$.

Then:

- for all $t^+ = 2$ indices i for which $t < \underline{x}_i$, we select $x_i = \overline{x}_i$; so, we select

  $$x_4 = \overline{x}_4 = e^3 \text{ and } x_5 = \overline{x}_5 = e^4;$$

- from the $n - t^+ - t^- = 5 - 2 - 0 = 3$ indices for which $\underline{x}_i \le t \le \overline{x}_i$, we select $k - t^+ = 0$ indices with the largest values of $\overline{x}_i$; since $k - t^+ = 0$, we skip this step;
- for one of the remaining indices, e.g., for the first of them $i = 1$, we take

  $$x_1 = t = e^1;$$

- for all other remaining indices, we take $x_i = \underline{x}_i$, i.e., we take

  $$x_2 = \underline{x}_2 = e^0 \text{ and } x_3 = \underline{x}_3 = e^1.$$

Then, we compute the value $H(x_1, \ldots, x_n)$ for the selected x_i. Here, $x_{(n-k)} = x_{(3)} = t = e^1$, and the larger values $x_{(4)}$ and $x_{(5)}$ are equal to e^3 and e^4. Thus,

$$H(x_1, \ldots, x_n) = \frac{1}{2} \cdot (\ln(x_{(4)}) + \ln(x_{(5)})) - \ln(x_{(3)}) =$$

$$\frac{1}{2} \cdot (3 + 4) - 1 = 3.5 - 1 = 2.5. \tag{49}$$

So, for $t = e^1$, we have $\overline{H}(t) = 2.5$.

Computing $\overline{H}$: case of $t = e^2$. After that, we consider the value $t = e^2$. In this case:

- we have $t^+ = 0$ intervals for which $t < \underline{x}_i$ and
- we have $t^- = 0$ intervals for which $\overline{x}_i < t$.

Then:

- from the $n - t^+ - t^- = 5 - 0 - 0 = 5$ indices for which $\underline{x}_i \le t \le \overline{x}_i$, we select $k - t^+ = 2 - 0 = 2$ indices with the largest values of $\overline{x}_i$; e.g., $i = 4$ and $i = 5$; for these values, we take $x_i = \overline{x}_i$, i.e., we take

$$x_4 = \overline{x}_4 = e^3 \text{ and } x_5 = \overline{x}_5 = e^4;$$

- for one of the remaining indices, e.g., for the first of them $i = 1$, we take

$$x_1 = t = e^2;$$

- for all other remaining indices, we take $x_i = \underline{x}_i$, i.e., we take

$$x_2 = \underline{x}_2 = e^0 \text{ and } x_3 = \underline{x}_3 = e^1.$$

Then, we compute the value $H(x_1, \ldots, x_n)$ for the selected x_i. Here, $x_{(n-k)} = x_{(3)} = t = e^2$, and the larger values $x_{(4)}$ and $x_{(5)}$ are equal to e^3 and e^4. Thus,

$$H(x_1, \ldots, x_n) = \frac{1}{2} \cdot (\ln(x_{(4)}) + \ln(x_{(5)})) - \ln(x_{(3)}) =$$

$$\frac{1}{2} \cdot (3 + 4) - 2 = 3.5 - 2 = 1.5. \tag{50}$$

So, for $t = e^2$, we have $\overline{H}(t) = 1.5$.

Computing $\overline{H}$: case of $t = e^3$. For $t = e^3$:

- we have $t^+ = 0$ intervals for which $t < \underline{x}_i$, and
- we have $t^- = 1$ interval for which $\overline{x}_i < t$: the interval

$$[\underline{x}_1, \overline{x}_1] = [e^1, e^2].$$

Then:

- from the $n - t^+ - t^- = 5 - 0 - 1 = 4$ indices for which $\underline{x}_i \le t \le \overline{x}_i$, we select $k - t^+ = 2 - 0 = 2$ indices with the largest values of $\overline{x}_i$; e.g., $i = 4$ and $i = 5$; for these values, we take $x_i = \overline{x}_i$, i.e., we take

$$x_4 = \overline{x}_4 = e^3 \text{ and } x_5 = \overline{x}_5 = e^4;$$

- for one of the remaining indices, e.g., for the first of them $i = 2$, we take

$$x_2 = t = e^3;$$

- for all other remaining indices, we take $x_i = \underline{x}_i$, i.e., we take

$$x_1 = \underline{x}_1 = e^1 \text{ and } x_3 = \underline{x}_3 = e^1.$$

Then, we compute the value $H(x_1, \ldots, x_n)$ for the selected x_i. Here, $x_{(n-k)} = x_{(3)} = t = e^3$, and the larger values $x_{(4)}$ and $x_{(5)}$ are equal to e^3 and e^4. Thus,

$$H(x_1, \ldots, x_n) = \frac{1}{2} \cdot (\ln(x_{(4)}) + \ln(x_{(5)})) - \ln(x_{(3)}) =$$

$$\frac{1}{2} \cdot (3 + 4) - 3 = 3.5 - 3 = 0.5. \tag{51}$$

So, for $t = e^3$, we have $\overline{H}(t) = 0.5$.

Computing $\overline{H}$: case of $t = e^4$. Finally, for $t = e^4$:

- we have $t^+ = 0$ intervals for which $t < \underline{x}_i$, and
- we have $t^- = 4$ intervals for which $\overline{x}_i < t$:

$$[\underline{x}_1, \overline{x}_1] = [e^1, e^2]; \quad [\underline{x}_2, \overline{x}_2] = [e^0, e^3]; \quad [\underline{x}_3, \overline{x}_3] = [e^1, e^3];$$

$$[\underline{x}_4, \overline{x}_4] = [e^2, e^3].$$

Since $t^- = 4 > n - k - 1 = 5 - 2 - 1 = 2$, we dismiss this endpoints.

Computing $\overline{H}$: final step. At the end, we take the largest of the three values $\overline{H}(t)$ as the desired value $\overline{H}$;

$$\overline{H} = \max(\overline{H}(e^1), \overline{H}(e^2), \overline{H}(e^3)) = \max(2.5, 1.5, 0.5) = 2.5. \tag{52}$$

Computing $\underline{H}$. Here, we also consider all the endpoints t – except for the values $t = e^0$ and $t = e^4$ which, as we already know from computing $\overline{H}$, will be dismissed.

Computing $\underline{H}$: case of $t = e^1$. First, we consider the value $t = e^1$. In this case:

- We have $t^+ = 2$ intervals for which $t < \underline{x}_i$: intervals

$$[\underline{x}_4, \overline{x}_4] = [e^2, e^3] \text{ and } [\underline{x}_5, \overline{x}_5] = [e^2, e^4].$$

- We also have $t^- = 0$ intervals for which $\overline{x}_i < t$.

Then:

- for all $t^+ = 2$ indices i for which $t < \underline{x}_i$, we select $x_i = \underline{x}_i$, i.e., we take

$$x_4 = \underline{x}_4 = e^2 \text{ and } x_5 = \underline{x}_5 = e^2;$$

- from the $n - t^+ - t^- = 5 - 2 - 0 = 3$ indices for which $\underline{x}_i \le t \le \overline{x}_i$, we select $k + 1 - t^+ = 2 + 1 - 2 = 1$ index for which we take $x_i = t$; for example, we select the smallest such index $i = 1$, and select

$$x_1 = t = e^2;$$

- for the remaining indices, we take $x_i = \underline{x}_i$, i.e., we take

$$x_2 = \underline{x}_2 = e^0 \text{ and } x_3 = \underline{x}_3 = e^1.$$

Then, we compute the value $H(x_1, \ldots, x_n)$ for the selected x_i. Here, $x_{(n-k)} = x_{(3)} = t = e^1$, and the larger values $x_{(4)}$ and $x_{(5)}$ are equal to e^2 and e^2. Thus,

$$H(x_1, \ldots, x_n) = \frac{1}{2} \cdot (\ln(x_{(4)}) + \ln(x_{(5)})) - \ln(x_{(3)}) =$$

$$\frac{1}{2} \cdot (2 + 2) - 1 = 2 - 1 = 1. \tag{53}$$

Computing $\underline{H}$: case of $t = e^2$. After that, we consider the value $t = e^2$. In this case:

- we have $t^+ = 0$ intervals for which $t < \underline{x}_i$ and
- we have $t^- = 0$ intervals for which $\overline{x}_i < t$.

Then:

- from the $n - t^+ - t^- = 5 - 0 - 0 = 5$ indices for which $\underline{x}_i \le t \le \overline{x}_i$, we select $k + 1 - t^+ = 2 + 1 - 0 = 3$ indices for which we take $x_i = t$; for example, we select the smallest such indices $i = 1$, $i = 2$, and $i = 3$, and select

$$x_1 = x_2 = x_3 = t = e^2;$$

- for the remaining indices, we take $x_i = \underline{x}_i$, i.e., we take

$$x_4 = \underline{x}_4 = e^2 \text{ and } x_5 = \underline{x}_5 = e^2.$$

Then, we compute the value $H(x_1, \ldots, x_n)$ for the selected x_i. Here, $x_{(n-k)} = x_{(3)} = t = e^2$, and the larger values $x_{(4)}$ and $x_{(5)}$ are equal to e^2 and e^2. Thus,

$$H(x_1, \ldots, x_n) = \frac{1}{2} \cdot (\ln(x_{(4)}) + \ln(x_{(5)})) - \ln(x_{(3)}) =$$

$$\frac{1}{2} \cdot (2 + 2) - 2 = 2 - 2 = 0. \tag{54}$$

Computing $\underline{H}$: case of $t = e^3$. Finally, we consider the value $t = e^3$. In this case:

- We have $t^+ = 0$ intervals for which $t < \underline{x}_i$, and
- we have $t^- = 1$ interval
$$[\underline{x}_1, \overline{x}_1] = [e^1, e^2]$$
for which $\overline{x}_i < t$.

Then:

- from the $n - t^+ - t^- = 5 - 0 - 1 = 4$ indices for which $\underline{x}_i \le t \le \overline{x}_i$, we select $k + 1 - t^+ = 2 + 1 - 0 = 3$ indices for which we take $x_i = t$; for example, we select the smallest such indices $i = 2$, $i = 3$, and $i = 4$, and select
$$x_2 = x_3 = x_4 = t = e^3;$$
- for the remaining indices, we take $x_i = \underline{x}_i$, i.e., we take
$$x_1 = \underline{x}_1 = e^1 \text{ and } x_5 = \underline{x}_5 = e^2.$$

Then, we compute the value $H(x_1, \ldots, x_n)$ for the selected x_i. Here, $x_{(n-k)} = x_{(3)} = t = e^3$, and the larger values $x_{(4)}$ and $x_{(5)}$ are equal to e^3 and e^3. Thus,

$$H(x_1, \ldots, x_n) = \frac{1}{2} \cdot (\ln(x_{(4)}) + \ln(x_{(5)})) - \ln(x_{(3)}) =$$

$$\frac{1}{2} \cdot (3 + 3) - 3 = 3 - 3 = 0. \tag{55}$$

Computing $\underline{H}$: final step. At the end, we take the smallest of the three values $\underline{H}(t)$ as the desired value $\underline{H}$;

$$\underline{H} = \max(\underline{H}(e^1), \underline{H}(e^2), \underline{H}(e^3)) = \max(1, 0, 0) = 0. \tag{56}$$

Result: for the given data, the range of possible values of the Hill estimator is equal to
$$[\underline{H}, \overline{H}] = [0, 2.5].$$

Comment. From the definition of the Hill estimator, it easily follows that this estimator is always non-negative. Indeed,

$$H = \frac{1}{k} \cdot \sum_{j=1}^{k} \ln(x_{(n-j+1)}) - \ln(x_{(n-k)}) = \frac{1}{k} \cdot \sum_{j=1}^{k} \ln(x_{(n-j+1)}) - \frac{1}{k} \cdot \sum_{j=1}^{k} \ln(x_{(n-k)}) =$$

$$\frac{1}{k} \cdot \sum_{j=1}^{k} (\ln(x_{(n-j+1)}) - \ln(x_{(n-k)})). \tag{57}$$

Thus, the smallest possible value $\underline{H}$ of the Hill estimator is also always non-negative: $\underline{H} \geq 0$.

One can also easily check that $H = 0$ if $k+1$ largest values x_i are equal to each other:

$$x_{(n-k)} = x_{(n+k+1)} = \ldots = x_{(n)}. \tag{58}$$

In this case, the smallest possible value $\underline{H}$ is ≤ 0, hence $\underline{H} = 0$.

In our example, we can have $x_1 = x_2 = \ldots = x_5 = e^2$, so it is indeed possible to have $H = 0$, thus $\underline{H} = 0$.

Acknowledgments

This work was supported in part by the National Science Foundation grants HRD-0734825 and DUE-0926721, by Grant 1 T36 GM078000-01 from the National Institutes of Health, by Grant MSM 6198898701 from MŠMT of Czech Republic, and by Grant 5015 "Application of fuzzy logic with operators in the knowledge based systems" from the Science and Technology Centre in Ukraine (STCU), funded by European Union.

This work was performed when M. Chiangpradit and W. Panichkitkosolkul were visiting the New Mexico State University. M. Chiangpradit was supported by the Thailand Commission on Higher Education Strategic Scholarships for Frontier Research Network program. W. Panichkitkosolkul was supported by a scholarship from Thailand Ministry of Science and Technology.

The authors are very thankful to Van Nam Huynh, Sa-aat Niwitpong, and Hung T. Nguyen for their encouragement and support, to all the participants of the International Symposium on Integrated Uncertainty Management and Applications IUM'10 (Japan Advanced Institute of Science and Technology JAIST, Ishikawa, Japan, April 7–12, 2010) for useful discussions.

References

[1] J. Aczel, *Lectures on Functional Equations and Their Applications*, Dover Publ., New York, 2006.

[2] L. Bachelier, *Théorie de la spéculation*, PhD Dissertation, l'Ecole Normal Supérieure, Paris, 1900.

[3] J. Beirlant, Y, Goegevuer, J. Teugels, and J. Segers, *Statistics of Extremes: Theory and Applications*, Wiley, Chichester, 2004.

[4] B. K. Chakrabarti, A. Chakraborti, and A. Chatterjee, *Econophysics and Sociophysics: Trends and Perspectives*, Wiley-VCH, Berlin, 2006.

[5] A. Chatterjee, S. Yarlagadda, B. K. Chakrabarti, *Econophysics of Wealth Distributions*, Springer-Verlag Italia, Milan, 2005.

[6] C. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, MIT Press, Boston, Massachusetts, 2009.

[7] J. D. Farmer and T. Lux (eds.), *Applications of statistical physics in economics and finance*, a special issue of the *Journal of Economic Dynamics and Control*, 2008, Vol. 32, No. 1, pp. 1–320.

[8] S. Ferson, L. Ginzburg, V. Kreinovich, L. Longpré, and M. Aviles, "Computing Variance for Interval Data is NP-Hard", *ACM SIGACT News*, 2002, Vol. 33, No. 2, pp. 108–118.

[9] S. Ferson, L. Ginzburg, V. Kreinovich, L. Longpré, and M. Aviles, "Exact Bounds on Finite Populations of Interval Data", *Reliable Computing*, 2005, Vol. 11, No. 3, pp. 207–233.

[10] X. Gabaix, G. Parameswaran, P. Vasiliki, and H. E. Stanley, "Understanding the cubic and half-cubic laws of financial fluctuations", *Physica A*, 2003, Vol. 324, pp. 1–5.

[11] X. Gabaix, G. Parameswaran, P. Vasiliki, and H. E. Stanley, "A theory of power-law distributions in financial market fluctuations", *Nature*, 2003, Vol. 423, No. 6937, pp. 267–270.

[12] B. M. Hill, "A simple approach to inference about the tail of the distribution", *Annals of Statistics*, 1975, Vol. 3, pp. 1163–1174.

[13] C. Hu, R. B. Kearfott, A. de Korvin, and V. Kreinovich (eds.), *Knowledge Processing with Interval and Soft Computing*, Springer Verlag, London, 2008.

[14] R. B. Kearfott and V. Kreinovich (eds.), *Applications of Interval Computations*, Kluwer, Dordrecht, 1996.

[15] V. Kreinovich, A. Lakeyev, J. Rohn, and P. Kahl, *Computational complexity and feasibility of data processing and interval computations*, Kluwer, Dordrecht, 1998.

[16] T. Magoč and V. Kreinovich, "Empirical Formulas for Economic Fluctuations: Towards A New Justification", *Proceedings of the 28th North American Fuzzy Information Processing Society Annual Conference NAFIPS'09*, Cincinnati, Ohio, June 14–17, 2009.

[17] B. Mandelbrot, "The variation of certain speculative prices", *J. Business*, 1963, Vol. 36, pp. 394–419.

[18] B. Mandelbrot, *The Fractal Geometry of Nature*, Freeman, San Francisco, California, 1983.

[19] B. Mandelbrot and R. L. Hudson, *The (Mis)behavior of Markets: A Fractal View of Financial Turbulence*, Basic Books, 2006.

[20] R. N. Mantegna and H. E. Stanley, *An Introduction to Econophysics: Correlations and Complexity in Finance*, Cambridge University Press, Cambridge, Massachusetts, 1999.

[21] N. Markovich (ed.), *Nonparametric Analysis of Univariate Heavy-Tailed Data: Research and Practice*, Wiley, Chichester, 2007.

[22] J. McCauley, *Dynamics of Markets, Econophysics and Finance*, Cambridge University Press, Cambridge, Massachusetts, 2004.

[23] R. E. Moore, R. B. Kearfott, and M. J. Cloud, *Introduction to Interval Analysis*, SIAM Press, Philadelphia, Pennsylvania, 2009.

[24] H. T. Nguyen and V. Kreinovich, *Applications of continuous mathematics to computer science*, Kluwer, Dordrecht, 1997.

[25] C. Papadimitriou, *Computational Complexity*, Addison Welsey, Reading, Massachusetts, 1994.

[26] J. Pexider, Notiz uber Funktionaltheoreme, Monatsch. Math. Phys., 1903, Vol. 14, pp. 293–301.

[27] S. Rabinovich, *Measurement Errors and Uncertainties: Theory and Practice*, Springer Verlag, New York, 2005.

[28] S. I. Resnick, *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*, Springer-Varlag, New York, 2007.

[29] B. Roehner, *Patterns of Speculation - A Study in Observational Econophysics*, Cambridge University Press, Cambridge, Massachusetts, 2002.

[30] D. G. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman & Hall/CRC Press, Boca Raton, Florida, 2007.

[31] H. E. Stanley, "Econophysics and the current economic turmoil", *American Physical Society News*, 2008, Vol. 17, No. 11, p. 8.

[32] H. E. Stanley, L. A. N. Amaral, P. Gopikrishnan, and V. Plerou, "Scale invariance and universality of economic fluctuations", *Physica A*, 2000, Vol. 283, pp. 31–41.

[33] P. Vasiliki and H. E. Stanley, "Stock return distributions: tests of scaling and universality from three distinct stock markets", *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics*, 2008, Vol. 77, No. 3, Pt. 2, Publ. 037101.

[34] I. Weissman, "Estimation of parameters and larger quantiles based on the k largest observations", *Journal of the American Statistical Association*, 1978, Vol. 73, No. 364, pp. 812–815.