Asymptotic properties of the Hill estimator for error contaminated data

Mihyun Kim

Joint work with Piotr Kokoszka

Department of Statistics
Colorado State University

EVA, July 4 2019
Outline

1. Motivation
2. Main results
3. Finite sample behavior
The Hill estimator

- Suppose that $X_1, \ldots, X_n$ are i.i.d. nonnegative random variables with common distribution $F_X$, which has **regularly varying** tail probabilities ($\bar{F}_X \in RV_{-\alpha}$):

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

- $\alpha^{-1}$ is estimated by **the Hill estimator**

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

with the convention that $X_{(1)}$ is the largest order statistic.
Motivation

- In many applications, data are contaminated by noise, measurement errors or roundoff errors.

- An example of internet traffic anomalies

**Fig 1:** A map showing 14 two-directional links of the backbone internet network in the United States known as Internet2.
Motivation - Internet Traffic Anomalies

Fig 2: Anomaly extraction process implemented by Bandara et al. (2014)

- Due to a huge amount of data to be processed, the algorithm computes an anomaly arrival time only with the precision of five minutes.
We assume that we observe $Y_i = X_i + \varepsilon_i$, $1 \leq i \leq n$, where $\bar{F}_X \in RV_{-\alpha}$, $\{\varepsilon_i\}$ are i.i.d. random errors following $F_\varepsilon$, and independent of the $\{X_i\}$.

For example, in the case of the internet traffic data, $Y_i = X_i + \varepsilon_i$ where
- the $Y_i$ are observations,
- the $X_i$ are “true” interarrival time,
- the errors $\varepsilon_i$ are distributed on $[-1, 1]$. (We use 5 minutes as a unit lag.)
Questions

- The Hill estimator for the observations $Y_i = X_i + \varepsilon_i$ is defined as

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

In our context, $\hat{H}_{k,n}$ is the estimator that can be actually used.

- The question is, what would be suitable assumptions on the measurement errors $\varepsilon_i$ to make $\hat{H}_{k,n}$ consistent or asymptotically normal?

Suppose that the $X_i$ are i.i.d. with $\bar{F}_X \in RV_{-\alpha}$. Then

$$H_{k,n} := \frac{1}{k} \sum_{i=1}^{k-1} \log \frac{X(i)}{X(k)} \overset{P}{\to} \frac{1}{\alpha},$$

as

$$n \to \infty, \quad k \to \infty, \quad \frac{k}{n} \to 0. \quad (1)$$
Consistency

Theorem 1

Suppose that the $X_i$ are i.i.d. random variables with $\bar{F}_X \in RV_{-\alpha}$, and $\varepsilon_i$ are i.i.d. $\sim \bar{F}_\varepsilon$, which satisfies

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

and independent of the $\{X_i\}$. Then any estimator of $\alpha$ computed from the $Y_i = X_i + \varepsilon_i$ is consistent as (1), if its counterpart computed from the unobservable $X_i$ is consistent.

- The measurement error $\varepsilon$ has a lighter tail than $X$; $\bar{F}_Y \in RV_{-\alpha}$ as well.

- This is no longer trivial if the $X_i$ are dependent. (New theorems are established)
Asymptotic normality

- To obtain the **asymptotic normality** of the Hill estimator centered by the exponent $\alpha^{-1}$, **second–order regular variation** needs to be assumed, see e.g. Hauesler and Teugels (1985, The Annals of Statistics), Resnick and Stărică (1997, Advances in Applied Probability), Resnick and Stărică (1997, Stochastic models).

- $2RV(-\alpha, \rho)$ ; there exists a positive function $g \in RV_{\rho}$ such that $g(t) \to 0$, as $t \to \infty$, and for $\alpha > 0$, $\rho \leq 0$, $K \neq 0$.

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

- Suppose that the $X_i$ are i.i.d. with $\bar{F}_X \in 2RV(-\alpha, \rho)$. Then

$$
\sqrt{k} \left( H_{k,n} - \frac{1}{\alpha} \right) \Rightarrow N(0, 1/\alpha^2),
$$

as

$$
\sqrt{k}g(b(n/k)) \to 0.
$$
Asymptotic normality - 2RV

- To obtain the centering $\alpha^{-1}$, we consider conditions on $F_X$; 2RV case and Pareto case.

**Theorem 2 - 2RV**

Suppose that $\bar{F}_X \in 2RV(\alpha, \rho)$, and $\varepsilon_i$ are i.i.d. $\sim \bar{F}_\varepsilon$, which satisfies

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

for some $\beta > \alpha - \rho$. In addition, the sequence $k = k(n)$ satisfies $\sqrt{k}g(b(n/k)) \to 0$ if $\rho > -1$ and $\sqrt{k}/b(n/k) \to 0$ if $\rho \leq -1$. Then, for $\alpha \geq 1$

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

- The measurement error $\varepsilon$ has a lighter tail than some power function. We need an additional restriction on the $k$. 

Mihyun Kim (CSU) Asymptotic properties of the Hill estimator for error contaminated data EVA, July 4 2019 11 / 20
Theorem 2 - Pareto

Suppose that $F_X(x) = x^{-\alpha}$, $x \geq 1$, and $\varepsilon_i$ are i.i.d. $\sim F_\varepsilon$, which satisfies

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

for some $\beta > \alpha + 1$. In addition, the sequence $k = k(n)$ satisfies $\sqrt{k/b(n/k)} \to 0$. Then, for $\alpha \geq 1$

$$\sqrt{k} \left( \hat{H}_{k,n} - \frac{1}{\alpha} \right) \Rightarrow N(0, 1/\alpha^2).$$
Finite sample behaviors

- Investigate **the impact of errors on the Hill estimator in finite samples**.

- We generate observations $Y_i = X_i + \varepsilon_i$, $i = 1, 2, \ldots, N$, $N = 500, 2000$.

- We use two models for the $X_i$, both having true tail index $\alpha = 2$.

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

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

- We consider four different distributions for the errors $\varepsilon_i$. For each of them, $P(|\varepsilon| > x) = o(x^{-\beta})$, for some $7 < \beta < 8$.

- a **normal** distribution with mean 0 and standard deviation $\sigma$
- a scaled $t$-distribution with 8 degrees of freedom (**scaled $t_8$**)
- a generalized Pareto distribution (**GPD**),
  
  $$P(|\varepsilon| > z) = (1 + \xi(z - \mu)/\sigma)^{-1/\xi},$$
  
  with location $\mu = 0$, shape $\xi = 1/8$, and scale $\sigma_{GPD}$.
- a **uniform** distribution on the interval $[-a, a]$, $a > 0$.

- For each model/error pair, we have 1000 replications.
The asymptotic level $1 - p$ confidence interval for $\alpha^{-1}$ implied by the Theorem 2 is
\[
\left( \frac{1}{\hat{\alpha}} - \frac{z_{p/2}}{\hat{\alpha} \sqrt{k}}, \frac{1}{\hat{\alpha}} + \frac{z_{p/2}}{\hat{\alpha} \sqrt{k}} \right),
\]
where $\hat{\alpha}^{-1} = \hat{H}_{k,n}$, and $z_q$ is the upper quantile of the standard normal distribution defined by $\Phi(z_q) = 1 - q$.

We investigate the impact of these errors on the empirical coverage probability of the interval.
Finite sample behaviors

- We examined four methods based on different underlying ideas of selecting a data-driven cut-off $k$.

  - **Hall**: It uses a bootstrap procedure to find the $k$ which minimizes the AMSE, introduced by Hall (1990, Journal of Multivariate Analysis).

  - **MAD**: It is based on minimizing a penalty function of the distance between the observed quantile and the fitted Pareto type tail. The mean absolute distance is used for the penalty and it is introduced by Danielsson et al. (2016).

  - **KS**: The underlying idea is the same, but the supremum of the absolute distance is used for the penalty. It is introduced by Danielsson et al. (2016).

  - **Eye**: It is an Eye–Ball technique trying to find a stable portion of the Hill plot and obtain the $k$ at which a considerable drop in the variance occurs, as $k$ increases. Danielsson et al. (2016).
<table>
<thead>
<tr>
<th>Method</th>
<th>Error Type</th>
<th>Error</th>
<th>Error SD/Model SD Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>0.01</td>
</tr>
<tr>
<td>Hall</td>
<td>Normal</td>
<td>88.9</td>
<td>87.6</td>
</tr>
<tr>
<td></td>
<td>scaled t₈</td>
<td>88.7</td>
<td>88.0</td>
</tr>
<tr>
<td></td>
<td>GPD</td>
<td>89.4</td>
<td>88.9</td>
</tr>
<tr>
<td></td>
<td>Uniform</td>
<td>89.1</td>
<td>88.2</td>
</tr>
<tr>
<td>MAD</td>
<td>Normal</td>
<td>97.0</td>
<td>97.4</td>
</tr>
<tr>
<td></td>
<td>scaled t₈</td>
<td>97.1</td>
<td>97.2</td>
</tr>
<tr>
<td>KS</td>
<td>Normal</td>
<td>83.4</td>
<td>82.2</td>
</tr>
<tr>
<td></td>
<td>scaled t₈</td>
<td>83.6</td>
<td>83.6</td>
</tr>
<tr>
<td>Eye</td>
<td>Normal</td>
<td>95.3</td>
<td>95.1</td>
</tr>
<tr>
<td></td>
<td>scaled t₈</td>
<td>95.3</td>
<td>95.4</td>
</tr>
</tbody>
</table>

**Table 1:** Proportion (in percent) of the approximate 95% confidence intervals including $1/\alpha$, for $n = 500$ and the **Pareto** model. The target coverage is 95%.
<table>
<thead>
<tr>
<th>Method</th>
<th>Error Type</th>
<th>Error SD</th>
<th>Model SD</th>
<th>Ratio</th>
<th>Error SD</th>
<th>Model SD</th>
<th>Ratio</th>
<th>Error SD</th>
<th>Model SD</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hall</td>
<td>Normal</td>
<td>75.3</td>
<td>75.6</td>
<td>75.0</td>
<td>12.9</td>
<td>8.8</td>
<td>37.2</td>
<td>34.7</td>
<td>75.8</td>
<td>75.3</td>
</tr>
<tr>
<td></td>
<td>scaled $t_8$</td>
<td>75.8</td>
<td>75.3</td>
<td>74.4</td>
<td>29.0</td>
<td>0.8</td>
<td>29.1</td>
<td>37.4</td>
<td>75.8</td>
<td>76.2</td>
</tr>
<tr>
<td></td>
<td>GPD</td>
<td>75.8</td>
<td>76.2</td>
<td>72.5</td>
<td>28.3</td>
<td>1.9</td>
<td>17.6</td>
<td>38.3</td>
<td>75.6</td>
<td>75.2</td>
</tr>
<tr>
<td></td>
<td>Uniform</td>
<td>75.6</td>
<td>75.2</td>
<td>74.0</td>
<td>33.8</td>
<td>26.4</td>
<td>35.0</td>
<td>30.3</td>
<td>76.6</td>
<td>76.2</td>
</tr>
<tr>
<td></td>
<td>scaled $t_8$</td>
<td>66.6</td>
<td>66.6</td>
<td>67.0</td>
<td>66.4</td>
<td>56.7</td>
<td>52.5</td>
<td>53.0</td>
<td>66.6</td>
<td>66.9</td>
</tr>
<tr>
<td></td>
<td>KS</td>
<td>93.6</td>
<td>93.9</td>
<td>93.4</td>
<td>93.7</td>
<td>92.6</td>
<td>88.7</td>
<td>77.8</td>
<td>93.6</td>
<td>93.9</td>
</tr>
<tr>
<td></td>
<td>scaled $t_8$</td>
<td>93.6</td>
<td>93.9</td>
<td>94.6</td>
<td>93.9</td>
<td>91.9</td>
<td>91.1</td>
<td>83.3</td>
<td>93.6</td>
<td>93.9</td>
</tr>
</tbody>
</table>

Table 2: Proportion (in percent) of the approximate 95% confidence intervals including $1/\alpha$, for $n = 500$ and the 2RV model. The target coverage is 95 percent.
Conclusions

- We derive broadly applicable **conditions on errors** under which the Hill estimator is **consistent** or **asymptotically normal**.

- From the simulation, **the impact of the errors is robust** for small ratios.

- There is no clear evidence that the coverage probability depends on the error distributions.
Thank you!