Extreme Value Analysis Without the Largest Values

Richard A. Davis[1]
Joint work with Gennady Samorodnitsky[2] and Jingjing Zou[1]

1. Department of Statistics, Columbia University; 2. School of Operations Research and Information Engineering, Cornell University

Motivation–heavy-tailed data

- Heavy-tails in data often modeled by a Pareto-like distribution, i.e.,

$$P(X > x) \sim \frac{1}{x^{\alpha}}, \quad x \geq 1, \tag{1}$$

where $\alpha > 0$.

Motivation–heavy-tailed data

- Heavy-tails in data often modeled by a Pareto-like distribution, i.e.,

$$P(X > x) \sim \frac{1}{x^{\alpha}}, \quad x \geq 1, \tag{1}$$

  where $\alpha > 0$.
- Goal is to estimate α.

Motivation–heavy-tailed data

- Heavy-tails in data often modeled by a Pareto-like distribution, i.e.,

$$P(X > x) \sim \frac{1}{x^{\alpha}}, \quad x \geq 1, \tag{1}$$

  where $\alpha > 0$.

- Goal is to estimate α.
- Equation (1) is only approximate for x large, i.e., $x > L$.

Motivation–heavy-tailed data

- Heavy-tails in data often modeled by a Pareto-like distribution, i.e.,

$$P(X > x) \sim \frac{1}{x^{\alpha}}, \quad x \geq 1, \tag{1}$$

  where $\alpha > 0$.
- Goal is to estimate α.
- Equation (1) is only approximate for x large, i.e., $x > L$.
- Use maximum likelihood estimation (gold standard) if (1) holds exactly.

Motivation–heavy-tailed data

- Heavy-tails in data often modeled by a Pareto-like distribution, i.e.,

$$P(X > x) \sim \frac{1}{x^{\alpha}}, \quad x \geq 1, \tag{1}$$

  where $\alpha > 0$.
- Goal is to estimate α.
- Equation (1) is only approximate for x large, i.e., $x > L$.
- Use maximum likelihood estimation (gold standard) if (1) holds exactly.
- What if equation is only approximate?

Hill Estimator

- Independent $X_1, X_2, \ldots, X_n \sim F(x)$, where F has *Pareto-like tails*.
- Tail index α
- Order statistics $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$
- Hill estimator for $1/\alpha$

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

Hill Plot (Without Truncation)

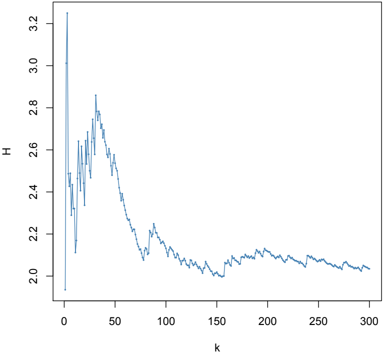


Figure: Hill plot of i.i.d. Pareto ($\alpha = 0.5$) variables ($n = 1000$)

Hill Plot

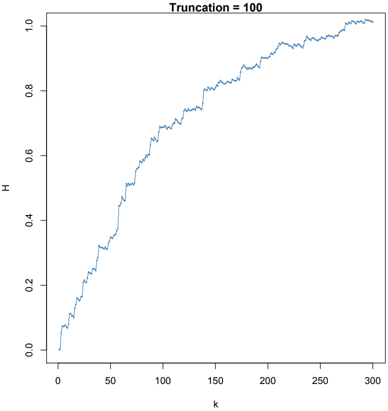


Figure: With 100 largest observations truncated

Example: Google+ Data

- A snapshot of the social network taken on Oct, 2012
- 76,438,791 nodes
- 1,442,504,499 edges

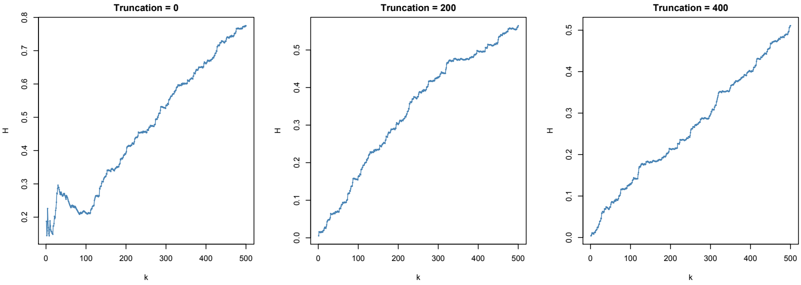


Figure: Hill Plots of In-degrees

Parametrization of Truncated Hill Estimator

- n: original sample size (before truncation)

Parametrization of Truncated Hill Estimator

- n: original sample size (before truncation)
- k_n: a sequence of integers such that $k_n \to \infty$ and $k_n/n \to 0$

Parametrization of Truncated Hill Estimator

- n: original sample size (before truncation)
- k_n: a sequence of integers such that $k_n \to \infty$ and $k_n/n \to 0$
- δk_n: observations truncated and NOT observed

Parametrization of Truncated Hill Estimator

- n: original sample size (before truncation)
- k_n: a sequence of integers such that $k_n \to \infty$ and $k_n/n \to 0$
- δk_n: observations truncated and NOT observed
- θk_n: number of top observations included in estimation

Parametrization of Truncated Hill Estimator

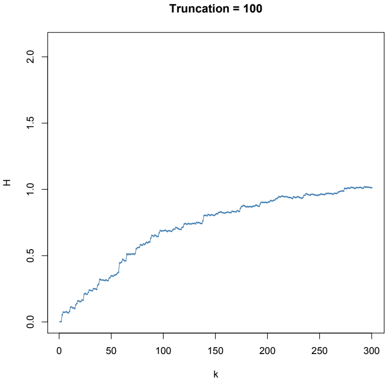
- n: original sample size (before truncation)
- k_n: a sequence of integers such that $k_n \to \infty$ and $k_n/n \to 0$
- δk_n: observations truncated and NOT observed
- θk_n: number of top observations included in estimation
- Truncated Hill estimator

$$H_n(\delta, \theta) = \frac{1}{\lfloor \theta k_n \rfloor} \sum_{i=1}^{\lfloor \theta k_n \rfloor} \log X_{(n - \lfloor \delta k_n \rfloor - i + 1)} - \log X_{(n - \lfloor \delta k_n \rfloor - \lfloor \theta k_n \rfloor)}$$

Functional Convergence of Truncated Hill Estimator

- $\sqrt{k_n}(H_n(\delta, \theta) - E(H_n))$ converges to a Gaussian process
- Different values of δ and α are distinguishable through behaviors of sample paths of H_n

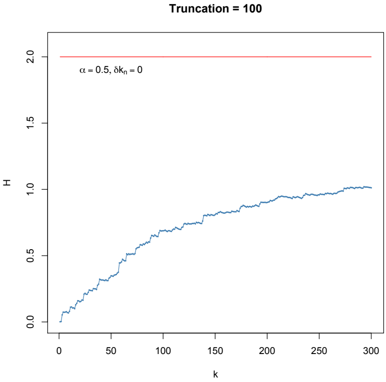
Figure: Pareto ($\alpha = 0.5$) variables ($n = 1000$, $k_n = 100$ and $\delta k_n = 100$)



Functional Convergence of Truncated Hill Estimator

- $\sqrt{k_n}(H_n(\delta, \theta) - E(H_n))$ converges to a Gaussian process
- Different values of δ and α are distinguishable through behavior of sample paths of H_n

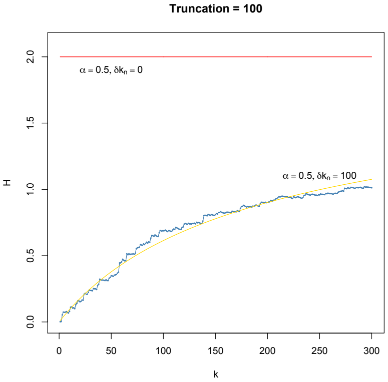
Figure: Pareto ($\alpha = 0.5$) variables ($n = 1000$, $k_n = 100$ and $\delta k_n = 100$)



Functional Convergence of Truncated Hill Estimator

- $\sqrt{k_n}(H_n(\delta,\theta) - E(H_n))$ converges to a Gaussian process
- Different values of δ and α are distinguishable through behaviors of sample paths of H_n

Figure: Pareto ($\alpha = 0.5$) variables ($n = 1000$, $k_n = 100$ and $\delta k_n = 100$)



Gaussian Processes

$n = 2000$ observations generated from Pareto distribution

- $k_n = 100$
- $\alpha = 0.5$

Figure: $\delta = 0$ (without truncation)

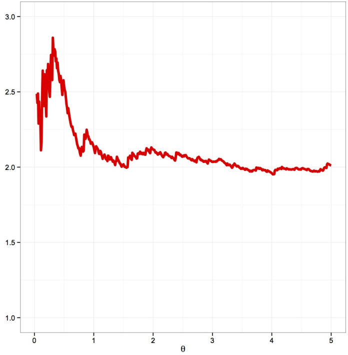
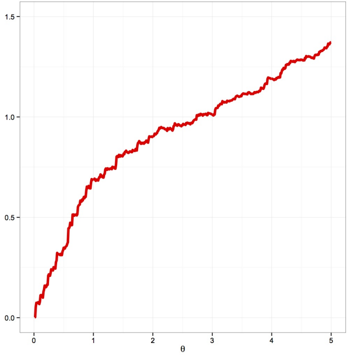


Figure: $\delta = 1$ (with truncation)



Gaussian Processes

Generate 50 sample paths from the limiting Gaussian processes

- $k_n = 100$
- $\alpha = 0.5$

Figure: $\delta = 0$ (without truncation)

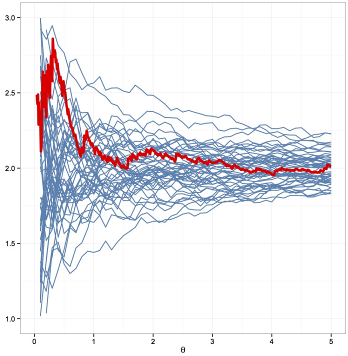
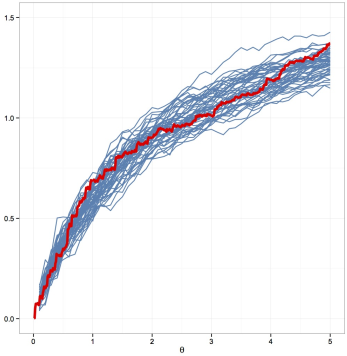


Figure: $\delta = 1$ (with truncation)



Functional Convergence of Truncated Hill Estimator

Theoretical conditions for the convergence

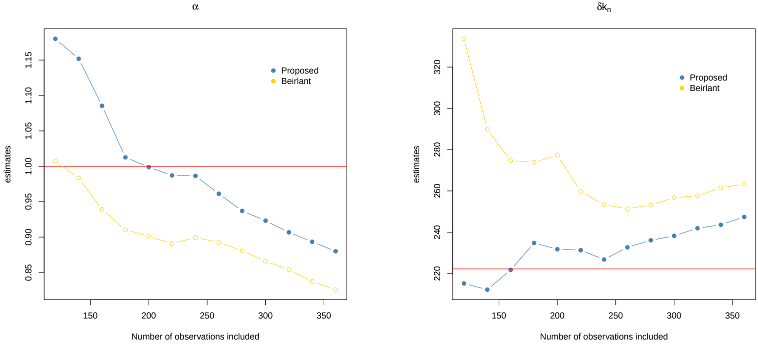
- F regularly varying
- Second-order regular variation condition
- Bias term in the mean of the Gaussian process if not Pareto

Estimation Procedure

- Estimate parameters based on the asymptotic joint distribution of $\{H_n\}$
- Solve for maximum likelihood estimators for
  - Number of truncated observations δk_n
  - Tail index α
- Beirlant et al. (2016) modeled truncation with threshold parameter T and estimated parameters based on Pareto likelihood

Estimation Results

- Cauchy distribution
- $\alpha = 1$, $n = 2000$, $k_n = 200$, truncation $\delta k_n = 200$
- Averaged estimation results of 200 independent simulations



Earthquake Data

- Earthquake fatalities by the U.S. Geological Survey (1900 - 2014) [1]
- $n = 125$ earthquakes with 1,000 or more deaths
- First apply the estimation procedures to the original data
- Then to the data with additional truncation of 10 top observations
- Estimations should reflect the truncation

[1] `http://earthquake.usgs.gov/earthquakes/world/world_deaths.php`

Earthquake Data

Figure: truncation = 0

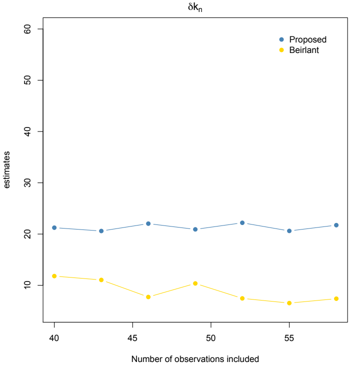


Figure: truncation = 10

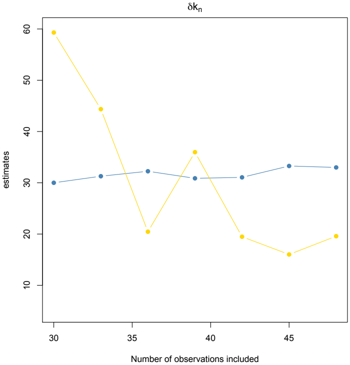


Figure: Estimates of number of truncation

Earthquake Data

Figure: truncation = 0

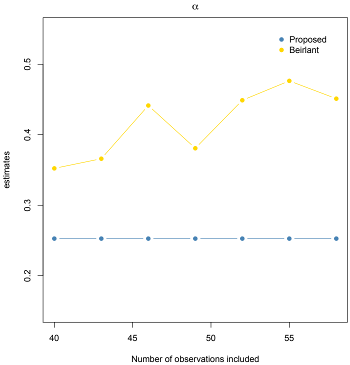


Figure: truncation = 10

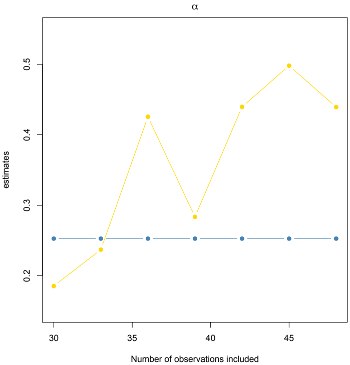


Figure: Estimates of the tail index α

Earthquake Data

Figure: truncation = 0

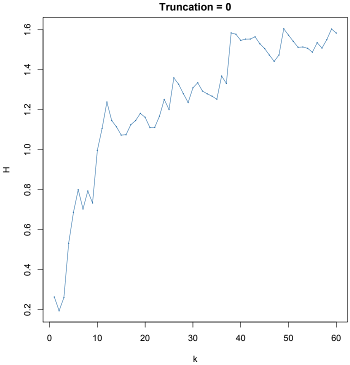


Figure: truncation = 10

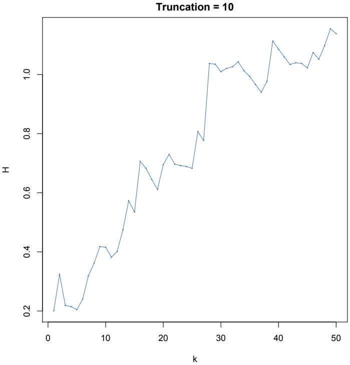


Figure: Hill estimators vs. fitted mean curves (with different number of observations included in estimation)

Earthquake Data

Figure: truncation = 0

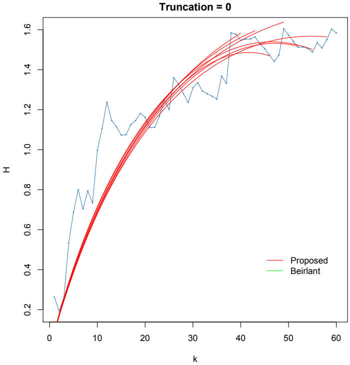


Figure: truncation = 10

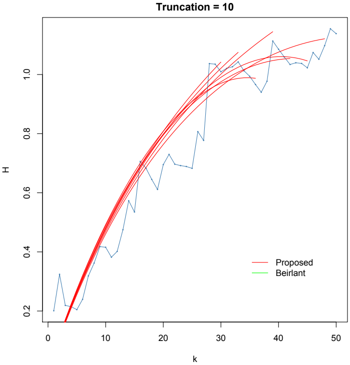


Figure: Hill estimators vs. fitted mean curves (with different number of observations included in estimation)

Earthquake Data

Figure: truncation = 0

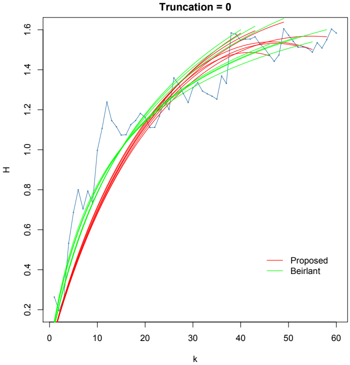


Figure: truncation = 10

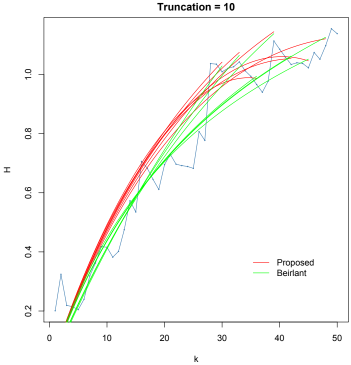


Figure: Hill estimators vs. fitted mean curves (with different number of observations included in estimation)

Google+ Data

Figure: truncation = 0

Figure: truncation = 400

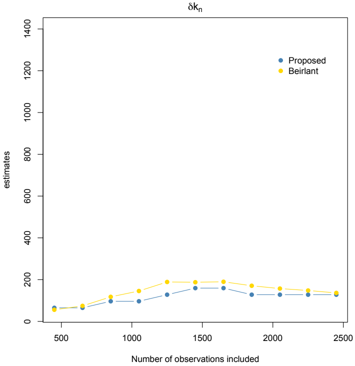


Figure: Estimates of number of truncation

Google+ Data

Figure: truncation = 0

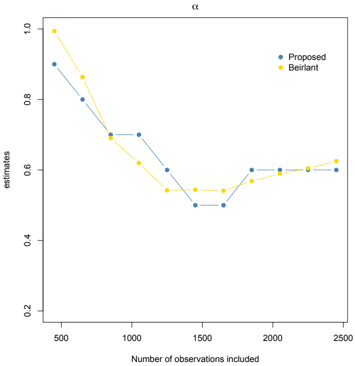


Figure: truncation = 400

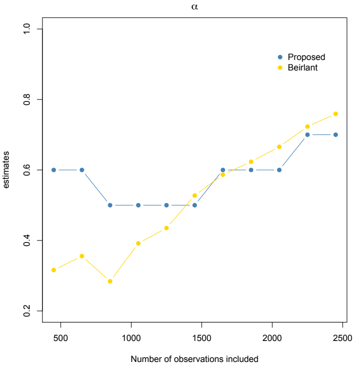


Figure: Estimates of tail index α

Google+ Data

Figure: truncation = 0

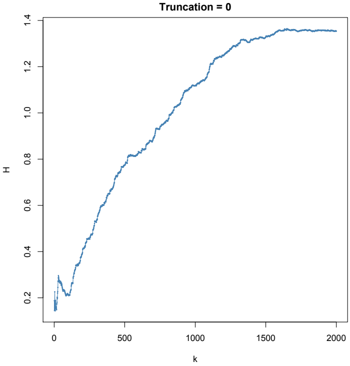


Figure: truncation = 400

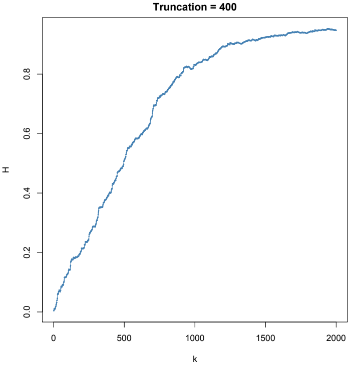


Figure: Hill estimators vs. fitted mean curves (with different number of observations included in estimation)

Google+ Data

Figure: truncation = 0

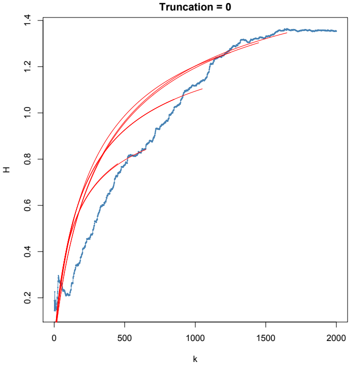


Figure: truncation = 400

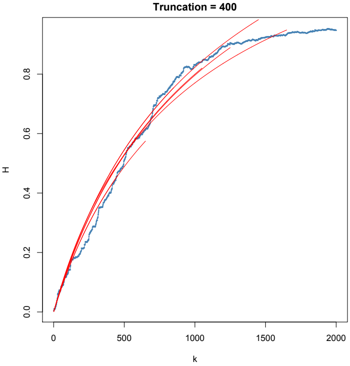


Figure: Hill estimators vs. fitted mean curves (with different number of observations included in estimation)

Google+ Data

Figure: truncation = 0

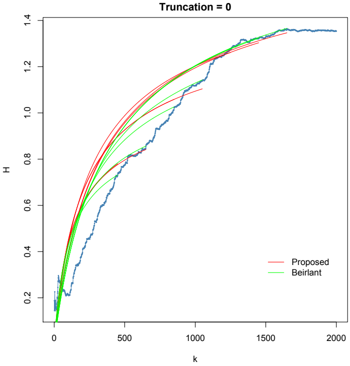


Figure: truncation = 400

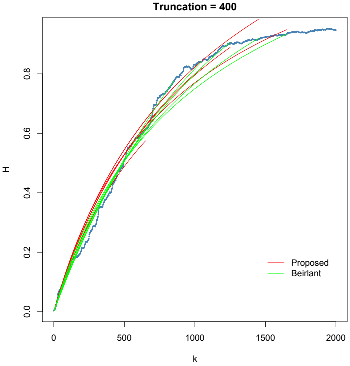


Figure: Hill estimators vs. fitted mean curves (with different number of observations included in estimation)

Reference

Jan Beirlant, Isabel Fraga Alves, and Ivette Gomes (2016). *Tail Fitting for truncated and non-truncated Pareto-type distributions*. **Extremes** 19.3, pp. 429–462. 26