Estimating heavy–tail exponents through max self–similarity

Stilian Stoev (sstoev@umich.edu)
Department of Statistics
University of Michigan

Probability & Statistics Colloquium
Michigan State University, October, 2006

Joint work with George Michailidis and Murad S. Taqqu

Part I: Motivation

Heavy tailed data

• A random variable X is said to be *heavy–tailed* if

$$\mathbb{P}\{|X| \geq x\} \sim L(x)x^{-\alpha}, \quad \text{as } x \to \infty,$$

for some $\alpha > 0$ and a slowly varying function L.

◦ Here we focus on the simpler but important context:

$$X \geq 0, \ \text{ a.s.} \quad \text{and} \quad \mathbb{P}\{X > x\} \sim Cx^{-\alpha}, \quad \text{as } x \to \infty.$$

◦ X (*infinite moments*) For $p > 0$,

$$\mathbb{E}X^p < \infty \quad \text{ if and only if } \quad p < \alpha.$$

In particular,

$$0 < \alpha \leq 2 \quad \Rightarrow \quad \text{Var}(X) = \infty$$

and

$$0 < \alpha \leq 1 \quad \Rightarrow \quad \mathbb{E}|X| = \infty.$$

• The estimation of the *heavy–tail exponent* α is an important problem with rich history.

• Why do we need heavy–tail models?

Every finite sample $X_1, \ldots, X_n$ has finite sample mean, variance and all sample moments!

Why consider heavy tailed models in practice?!

Why use heavy–tailed models?

"All models are wrong, but some are useful."

George Box

Let F and G be any two distributions with positive densities on $(0, \infty)$.

Let $\epsilon > 0$ and $x_1, \ldots, x_n \in (0, \infty)$ be arbitrary, then both:

$$\mathbb{P}_F\{X_i \in (x_i - \epsilon, x_i + \epsilon),\ i = 1, \ldots, n\} > 0$$

and

$$\mathbb{P}_G\{X_i \in (x_i - \epsilon, x_i + \epsilon),\ i = 1, \ldots, n\} > 0$$

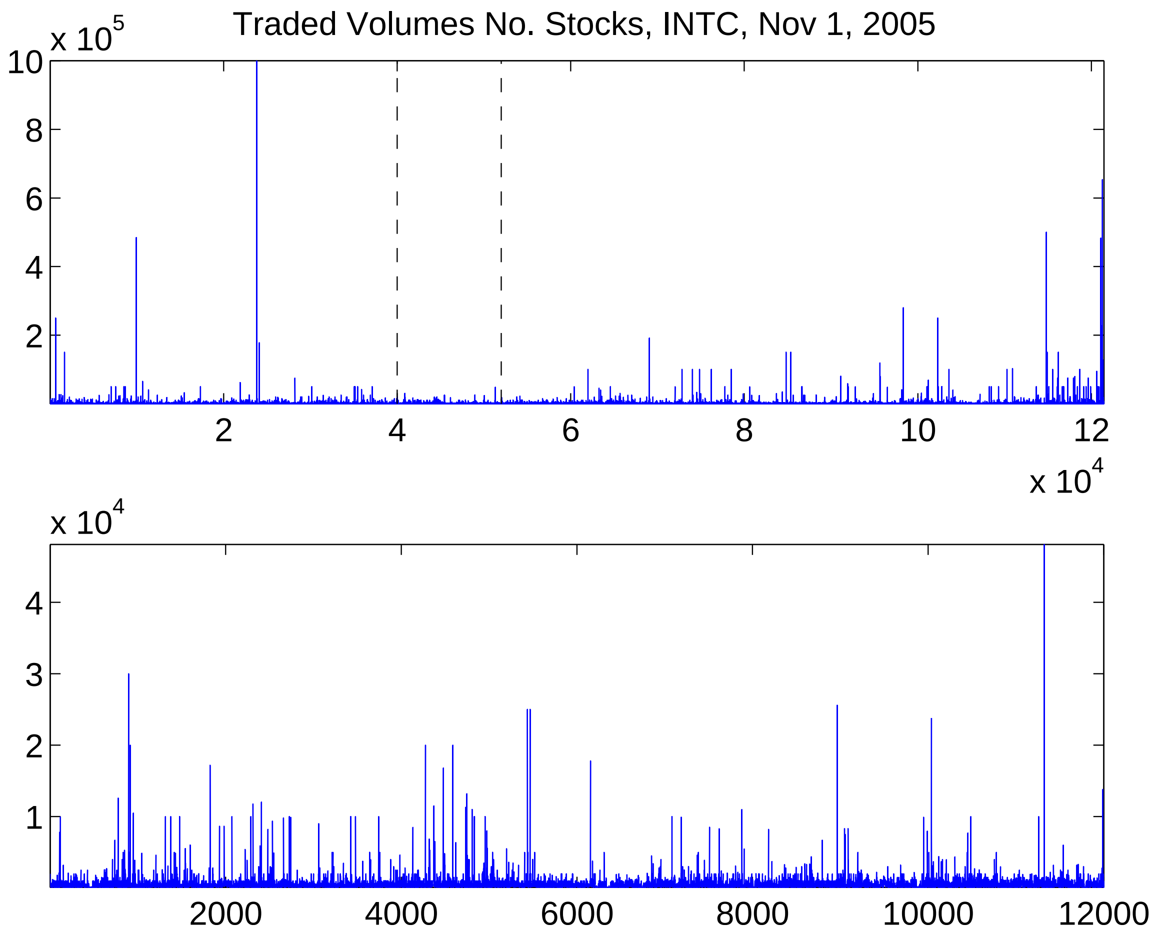
are positive!

- For a given sample, very many models apply.

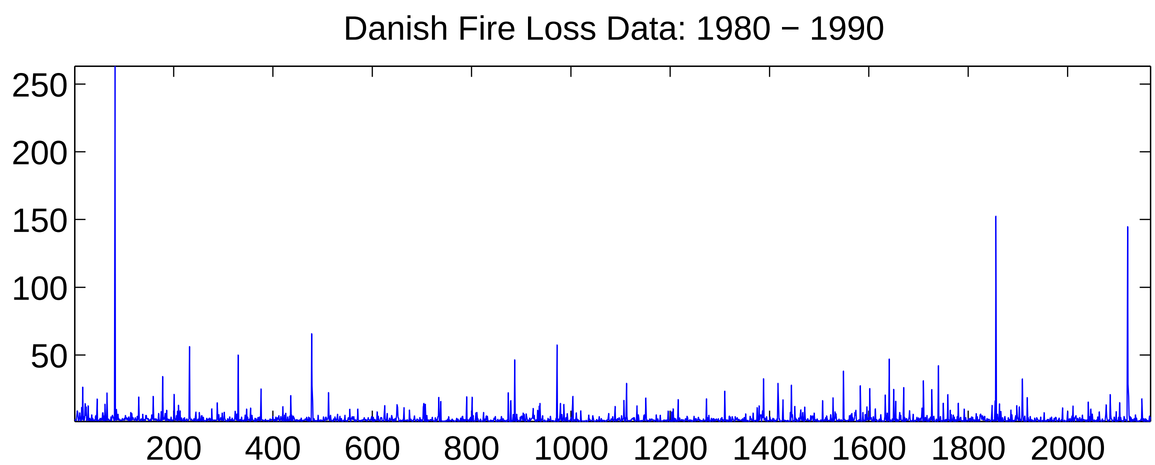
- The ones that continue to work as the sample grows are most suitable.

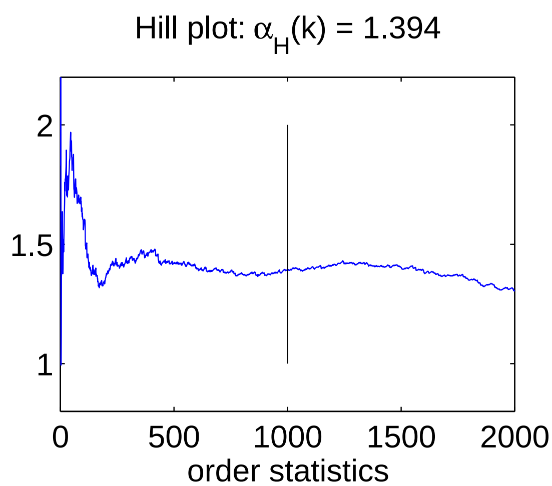
We next present real data sets of Financial, Insurance and Internet data. They can be *very heavy tailed*.

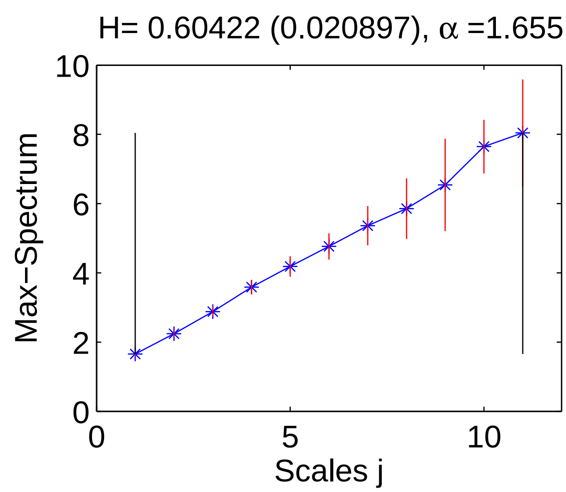
Traded volumes on the Intel stock



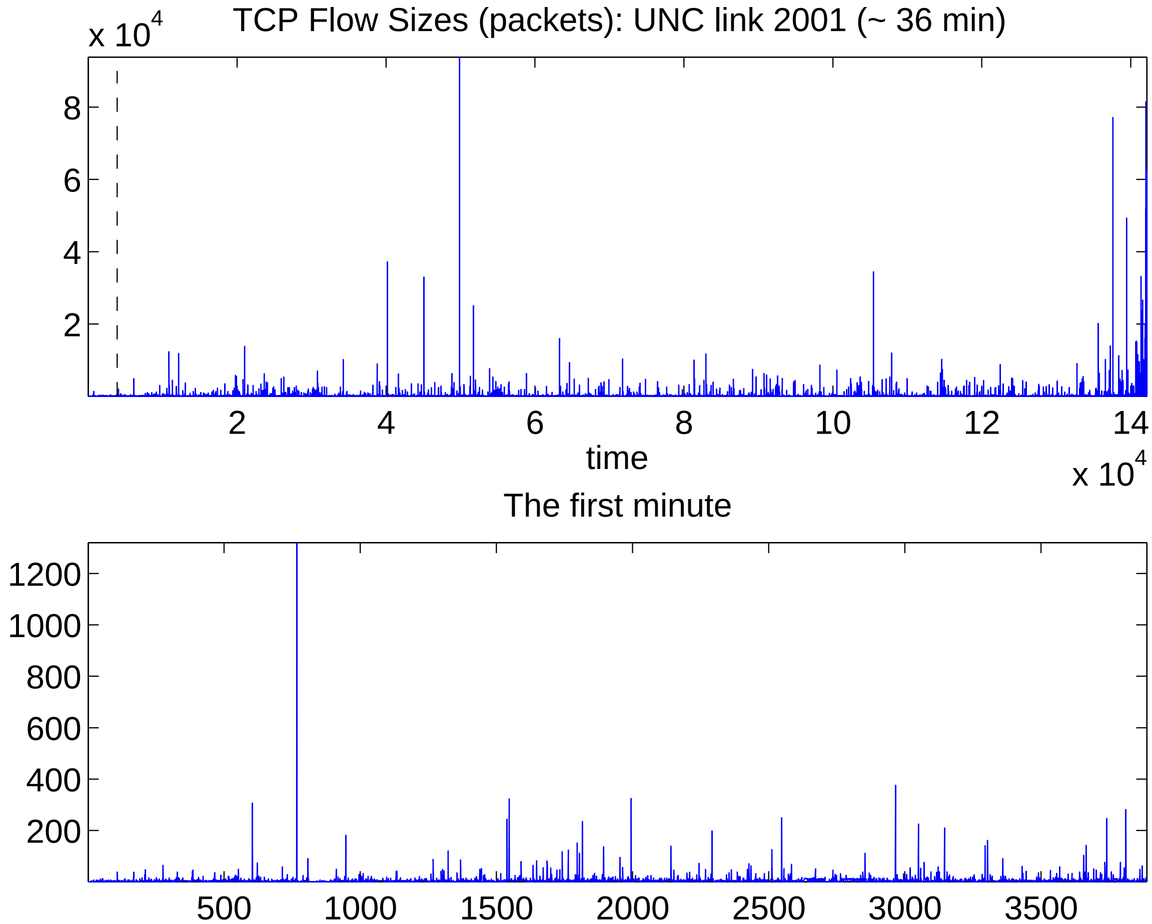
Insurance claims due to fire loss







TCP flow sizes (in number of packets)



History

• Hill (1975) worked out the MLE in the Pareto model $\mathbb{P}\{X > x\} = x^{-\alpha},\ x \geq 1$ and introduced the *Hill plot*:

$$\widehat{\alpha}_H(k) := (\frac{1}{k}\sum_{i=1}^{k} \log(X_{i,n}) - \log(X_{k+1,n}))^{-1},$$

where $X_{1,n} \geq X_{2,n} \geq \cdots \geq X_{k,n}$ are the *top–K order statistics* of the sample.

• How to choose k?

∘ pick k where the plot of $\widehat{\alpha}_H(k)$ vs. k stabilizes.

∘ serious problems in practice:

volatile & hard to interpret: "Hill horror plot"

confidence intervals

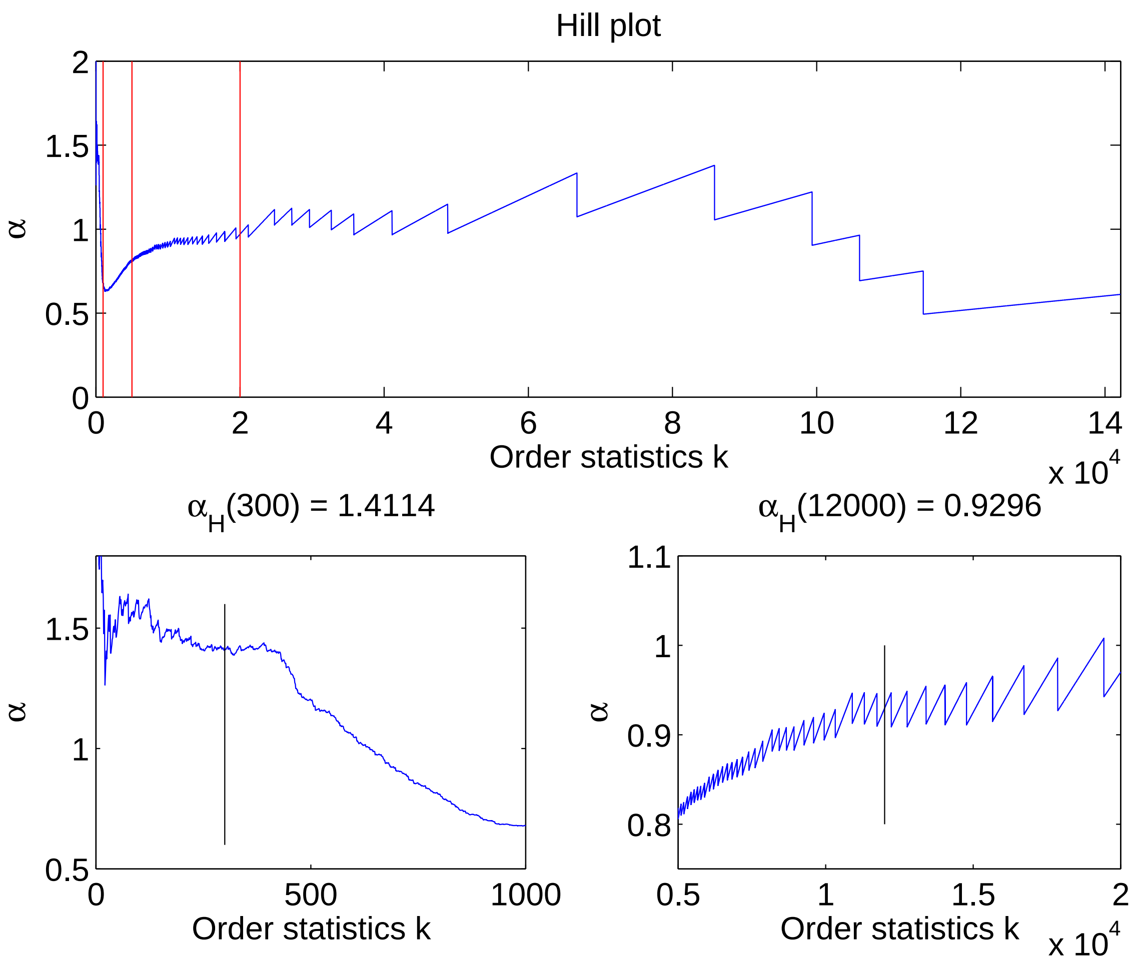
robustness

• Consistency and asymptotic normality resolved: Weissman (1978), Hall (1982) in semi–parametric setting.

• Many other estimators: *kernel based* Csörgő, Deheuvels and Mason (1985), *moment* Dekkers, Einmahl and de Haan (1989), among many others.

• Most estimators exploit the *top order statistics*.

Hill horror plots: TCP flow sizes



Part II: Max self–similarity & max–spectrum

Fréchet max–stable laws

Consider i.i.d. X_i's with $\mathbb{P}\{X_1 > x\} \sim Cx^{-\alpha},\ x \to \infty$. By the Fisher–Tippett–Gnedenko theorem:

$$\frac{1}{n^{1/\alpha}} \max_{1\le i\le n} X_i \equiv \frac{1}{n^{1/\alpha}} \bigvee_{i=1}^{n} X_i \stackrel{d}{\longrightarrow} Z, \quad \text{as } n \to \infty,$$

where Z is $\alpha-$Fréchet *extreme value distribution*:

$$\mathbb{P}\{Z \le x\} = \exp\{-Cx^{-\alpha}\}, \quad x > 0.$$

• The extreme value distributions are *max–stable*. In particular, for i.i.d. $\alpha-$Fréchet $Z,\&\ Z_i$'s:

$$Z_1 \vee \cdots \vee Z_n \stackrel{d}{=} n^{1/\alpha} Z.$$

• A time series of i.i.d. $\alpha-$Fréchet $\{Z_k\}$ is $1/\alpha-$max-self-similar:

$$\forall m \in \mathbb{N} : \{\vee_{1\le i\le m} Z_{m(k-1)+i}\}_{k\in\mathbb{N}} \stackrel{d}{=} m^{1/\alpha}\{Z_k\}_{k\in\mathbb{N}}.$$

∘ Block–maxima of size m have the same distribution as the original data, rescaled by the factor $m^{1/\alpha}$.

• Any heavy–tailed $\{X_k\}$ (i.i.d.) data set is *asymptotically max self–similar*.

Max–spectrum

Given a positive sample $X_1, \ldots, X_n$, consider the *dyadic* block–maxima:

$$D(j,k) := \max_{1 \le i \le 2^j} X_{2^j(k-1)+i} \equiv \bigvee_{i=1}^{2^j} X_{2^j(k-1)+i},$$

with $k = 1, \ldots, n_j = [n/2^j]$.

• In view of the *asymptotic scaling*:

$$\frac{1}{2^{j/\alpha}} D(j,k) \xrightarrow{d} Z, \quad (j \to \infty),$$

observe that

$$Y_j := \frac{1}{n_j} \sum_{j=1}^{n_j} \log_2 D(j,k) \simeq j/\alpha + c, \quad (j,\ n_j \to \infty)$$

where $c := \mathbb{E} \log_2 Z$.

∘ The last asymptotics "follow" from the LLN since $D(j,k)$'s are independent in k.

• The *max–spectrum* of the data is defined as the statistics:

$$Y_j,\ j = 1, \ldots, [\log_2(n)].$$

∘ Can identify α from the *slope* of the Y_j's vs. j, *for large* j's.

Max–spectrum estimators of α

Given the *max–spectrum* $Y_j,\ j = 1, \ldots, [\log_2(n)]$, define

$$\widehat{\alpha}(j_1, j_2) := (\sum_{j=j_1}^{j_2} w_j Y_j)^{-1},$$

where $\sum_j j w_j = 1$ and $\sum_j w_j = 0$.

• That is, use *linear regression* to estimate the slope $1/\alpha$.

• Issues:

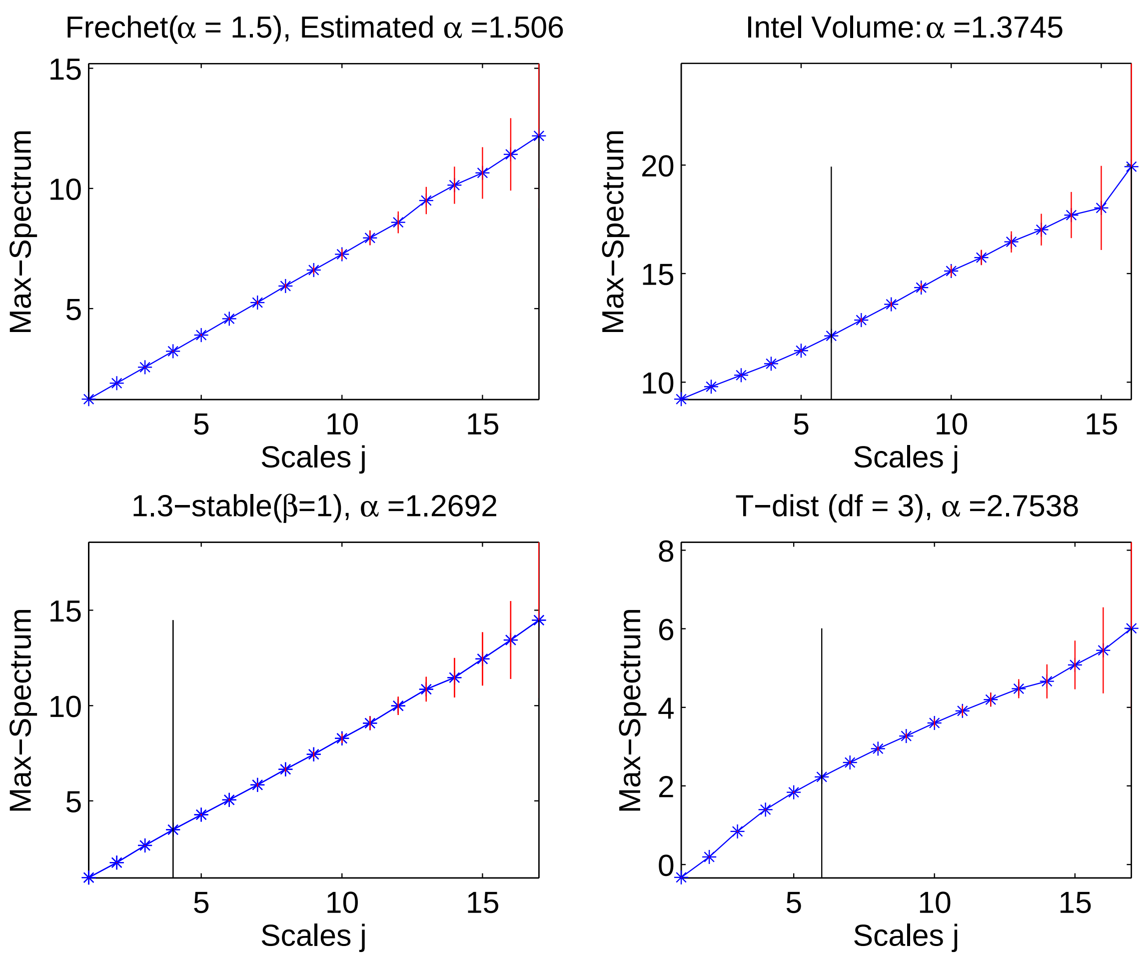
∘*weighted* or *generalized least squares* must be used, since

$$\text{Var}(Y_j) \propto 1/n_j \propto 2^j,$$

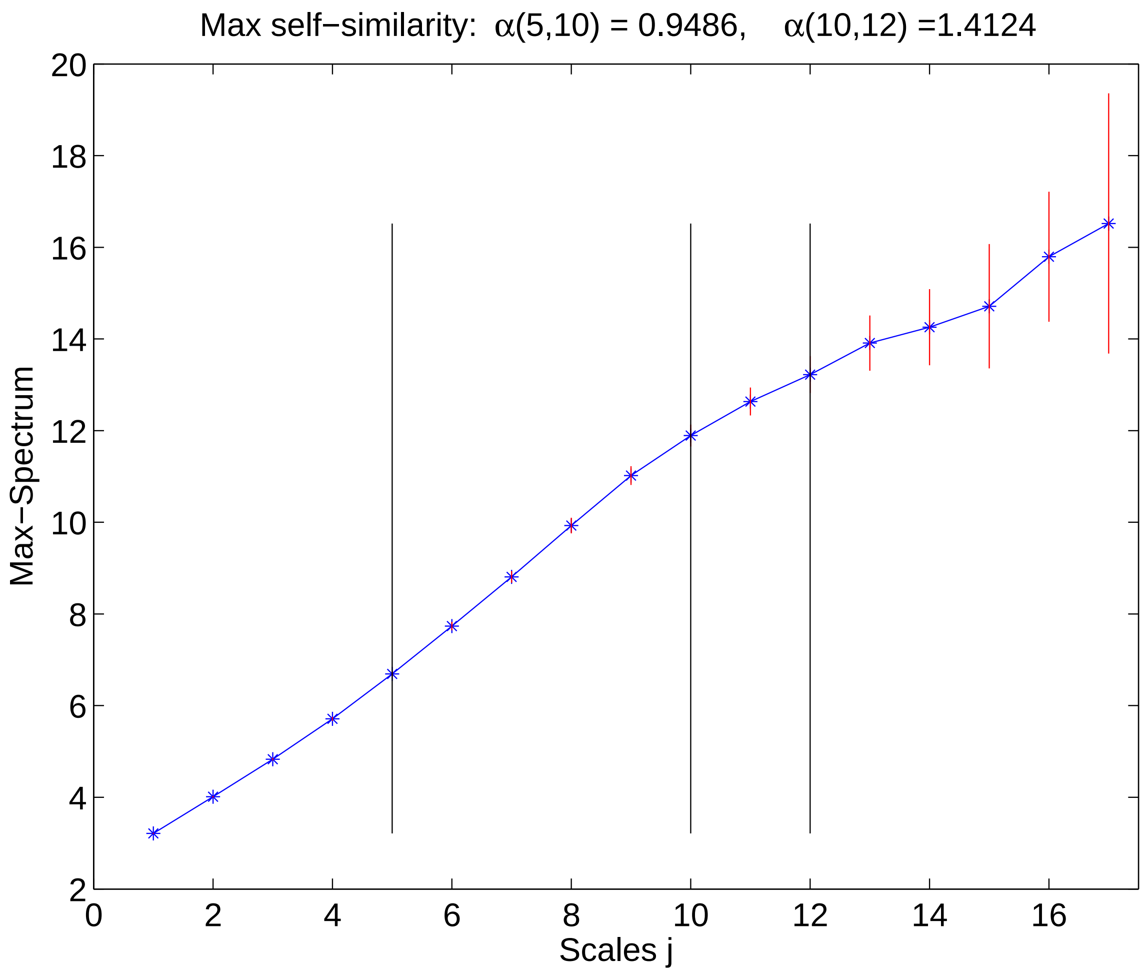
∘ The choice of the *scales* j_1 and j_2 is critical.

• These issues and confidence intervals, are addressed in Stoev, Michailidis and Taqqu (2006).

Examples of max–spectra

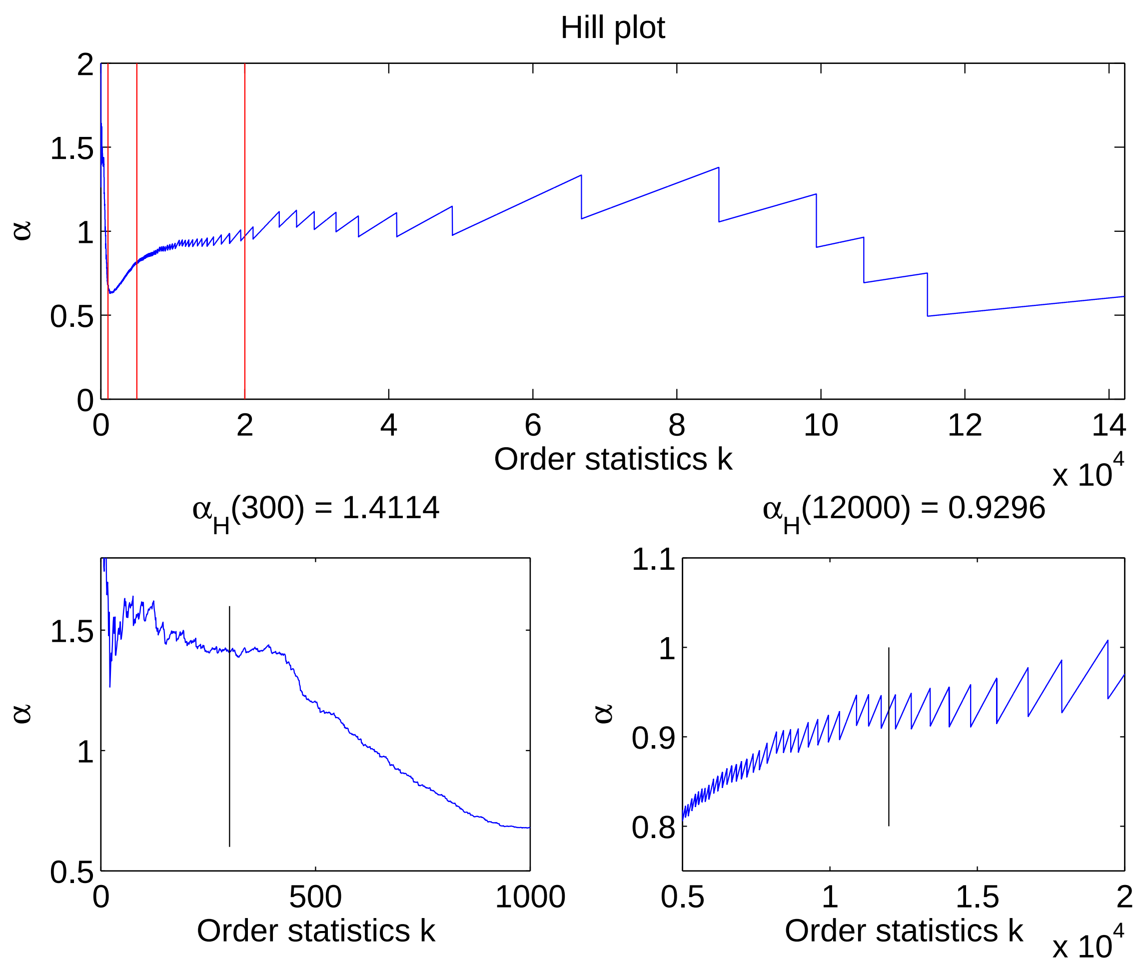


Robustness of the max–spectrum



Compare with the Hill horror plot of the Internet flow–size data (**next slide)**.

Hill horror plots: TCP flow sizes



Compare with the max–spectrum plot of the Internet flow–size data (**previous slide)**.

Part III: Asymptotic results

Asymptotic normality of the max–spectrum

Let $\mathbb{P}\{X_k \le x\} =: F(x)$, where

$$1 - F(x) = Cx^{-\alpha}(1 + \mathcal{O}(x^{-\beta})), \quad (x \to \infty), \qquad (1)$$

with some $\beta > 0$.

• Consider the max–spectrum $\{Y_j\}$ for the range of scales:

$$r(n) + j_1 \le j \le r(n) + j_2,$$

with *fixed* $j_1 < j_2$ and $r(n) \to \infty, \ n \to \infty$.

Theorem Under (1) and another technical condition,

$$\sup_{x\in\mathbb{R}} \Big| \mathbb{P}\{\sqrt{n_{j_2+r}}((\vec{\theta}, \vec{Y}_r) - (\vec{\theta}, \vec{\mu}_r)) \le x\} - \Phi(x/\sigma_{\vec{\theta}}) \Big|$$
$$\le C_{\vec{\theta}}(1/2^{r\beta/\alpha} + r2^{r/2}/\sqrt{n}), \qquad (2)$$

where Φ is the standard Normal c.d.f.

Here $\vec{Y}_r = (Y_{r+j})_{j=j_1}^{j_2}$ and

$$(\vec{\theta}, \vec{Y}_r) = \sum_{j=j_1}^{j_2} \theta_j Y_{r+j}, \qquad \mu_r(j) = (r+j)/\alpha + c$$

and

$$\sigma^2_{\vec{\theta}} = (\vec{\theta}, \Sigma_\alpha \vec{\theta}) := \sum_{i,j=j_1}^{j_2} \theta_i \Sigma_\alpha(i,j)\theta_j > 0.$$

The covariance structure of the max–spectrum

• The asymptotic covariance matrix Σ is the same as if the data were i.i.d. $\alpha-$Fréchet!

◦ The intuition is that the block–maxima

$$\{D(r+j,k),\ k=1,\dots,n_{r+j}\}$$

behave like i.i.d. $\alpha-$Fréchet variables, as $r=r(n)\to\infty$.

• The covariance entries are given by:

$$\Sigma_\alpha(i,j)=\alpha^{-2}2^{i\vee j-j_2}\psi(|i-j|),$$

with

$$\psi(a):=\mathrm{Cov}(\log_2(Z_1),\log_2(Z_1\vee(2^a-1)Z_2),\quad a\geq 0,$$

for i.i.d. $1-$Fréchet Z_1 & Z_2.

◦ Note that α appears only as a factor in Σ_α:

$$\Sigma_\alpha=\alpha^{-2}\Sigma_1.$$

It does not affect the *correlation structure* of the max–spectrum.

• GLS estimators for α use the matrix Σ_α.

• Asymptotic normality for $\widehat{\alpha}(r+j_1,r+j_2)$ follows from the last theorem.

• Stoev, Michailidis and Taqqu (2006) has details on *confidence intervals* for α and *automatic selection* of j_1&j_2.

Rates for moment functionals of maxima

Let $X_1, \ldots, X_n$ be i.i.d. from F and recall that

$$M_n := \frac{1}{n^{1/\alpha}} \bigvee_{1 \le i \le n} X_i \xrightarrow{d} Z, \quad (n \to \infty).$$

- Are there results on the rate of convergence

$$\mathbb{E}f(M_n) \longrightarrow \mathbb{E}f(Z), \quad (n \to \infty)$$

for "reasonable" f's?

∘ Pickands (1975) shows only the convergence of moments (no rates).

- Our approach: consider

$$F(x) = \exp\{-\sigma^\alpha(x) x^{-\alpha}\}, \quad x \in \mathbb{R}, \tag{3}$$

with

$$\sigma^\alpha(x) \longrightarrow C, \quad (x \to \infty).$$

∘ Note: $1 - F(x) \sim C x^{-\alpha}$, $x \to \infty$ is equivalent to (3).

- Extra assumption:

$$|\sigma^\alpha(x) - C| \le D x^{-\beta}, \quad \text{for large } x.$$

Rates (cont'd)

But (3) is very convenient to handle rates!

- Note that

$$\begin{aligned}\mathbb{P}\{M_n \le x\} &= \mathbb{P}\{X_1 \le n^{1/\alpha}x\}^n = F(n^{1/\alpha}x)^n \\ &= \exp\{-\sigma^\alpha(n^{1/\alpha}x)x^{-\alpha}\}. \qquad (4)\end{aligned}$$

◦ Thus,

$$\mathbb{E}f(M_n) = \int_0^\infty f(x)dF_n(x) = \int_0^\infty f(x)d\exp\{-\sigma^\alpha(n^{1/\alpha}x)x^{-\alpha}\},$$

and also

$$\mathbb{E}f(Z) = \int_0^\infty f(x)dG(x) = \int_0^\infty f(x)d\exp\{-Cx^{-\alpha}\}.$$

- Now, *integration by parts* yields:

$$\mathbb{E}(f(M_n) - f(Z)) = \int_0^\infty (G(x) - F_n(x))f'(x)dx.$$

◦ However $G(x)$ and $F_n(x)$ are of the same "exponential" form!

Rates (cont'd)

By the mean value theorem:

$$\begin{aligned} |F_n(x) - G(x)| &= |\exp\{-\sigma^\alpha(n^{1/\alpha}x)x^{-\alpha}\} - \exp\{-Cx^{-\alpha}\}| \\ &\le |\sigma^\alpha(n^{1/\alpha}x) - C|x^{-\alpha}\exp\{-cx^{-\alpha}\} \\ &\le Dn^{-\beta/\alpha}x^{-(\alpha+\beta)}\exp\{-cx^{-\alpha}\}, \quad \text{as } x \to \infty, \end{aligned}$$

• Thus, for any $\epsilon > 0$, as $n \to \infty$,

$$\int_\epsilon^\infty |F_n(x) - G(x)||f'(x)|dx$$
$$\le Dn^{-\beta/\alpha}\int_0^\infty |f'(x)|x^{-(\alpha+\beta)}e^{-cx^{-\alpha}}dx = \mathcal{O}(n^{-\beta/\alpha}).$$

• By taking $\epsilon = \epsilon_n \to 0$, and using a mild technical condition we can also bound the integral near zero

$$\int_0^{\epsilon_n} |F_n(x) - G(x)||f'(x)|dx.$$

Proposition If $\sigma^\alpha(x) \sim Dx^{-\beta}$, as $x \to \infty$, then as $n \to \infty$,

$$n^{\beta/\alpha}(\mathbb{E}f(M_n) - \mathbb{E}f(Z)) \longrightarrow D\int_0^\infty x^{-(\alpha+\beta)}f'(x)e^{-Cx^{-\alpha}}dx,$$

provided mild technical conditions on $\sigma(x)$ at 0 and on f at 0 and ∞ hold.

Rates (cont'd)

• We have thus obtained *exact rates* for *moment functionals*

$$\mathbb{E}f(\frac{1}{n^{1/\alpha}}\max_{1\le i\le n} X_i)$$

∘ They are valid for a large class of *absolutely continuous* f, including:

$$f(x) = \log_2(x), \quad \text{and} \quad f(x) = x^p,\ p \in (0,\alpha),$$

for example.

∘ More details can be found in Stoev, Michailidis and Taqqu (2006).

• As a corollary, we also get rates of convergence for

$$\mathrm{Cov}(\log_2 D(r+j_1,k), \log_2 D(r+j_2,k)), \quad \text{as } r = r(n) \to \infty.$$

• These tools and the Berry–Esseen Theorem, yield the *uniform asymptotic normality* of the *max–spectrum*.

Thank you!

References

Csörgő, S., Deheuvels, P. & Mason, D. (1985), 'Kernel estimates of the tail index of a distribution', *Annals of Statistics* **13**(3), 1050–1077.

Dekkers, A., Einmahl, J. & de Haan, L. (1989), 'A moment estimator for the index of an extreme–value distribution', *Ann. Statist.* **17**(4), 1833–1855.

Hall, P. (1982), 'On some simple estimates of an exponent of regular variation', *J. Roy. Stat. Assoc.* **44**, 37–42. Series B.

Hill, B. M. (1975), 'A simple general approach to inference about the tail of a distribution', *The Annals of Statistics* **3**, 1163–1174.

Pickands, J. (1975), 'Statistical inference using extreme order statistics', *Ann. Statist.* **3**, 119–131.

Stoev, S., Michailidis, G. & Taqqu, M. (2006), Estimating heavy–tail exponents through max self–similarity, Preprint.

Weissman, I. (1978), 'Estimation of parameters and large quantiles based on the k largest observations', *Journal of the American Statistical Association* **73**, 812–815.