Tail index estimation

S. Boucheron[1] and M.Thomas

[1]Laboratoire de Probabilités et Modèles Aléatoires
Université Paris-Diderot
DMA ENS

Math. Found. Learning Theory,
Dagstuhl 03/09/2015

Motivation(s)

Goals of EVT

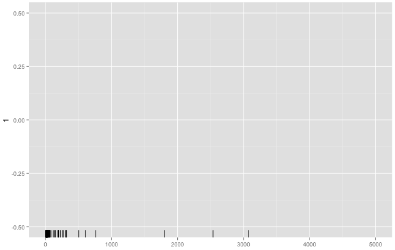
Extreme Value Theory (EVT) makes inferences about properties of the sampling distribution that lie outside the sample:

- inference of high order quantiles
- inference about probabilities of exceeding a threshold that is larger than the sample maximum

Goals of EVT

Extreme Value Theory (EVT) makes inferences about properties of the sampling distribution that lie outside the sample:

- inference of high order quantiles
- inference about probabilities of exceeding a threshold that is larger than the sample maximum

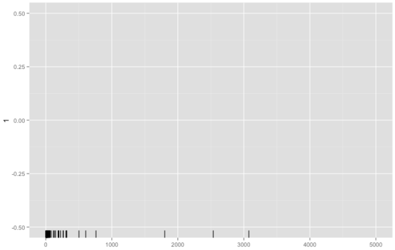


In order to face the challenge with reasonable chances of success, EVT makes assumption on the sampling distribution.

Goals of EVT

Extreme Value Theory (EVT) makes inferences about properties of the sampling distribution that lie outside the sample:

- ▹ inference of high order quantiles
- ▹ inference about probabilities of exceeding a threshold that is larger than the sample maximum



In order to face the challenge with reasonable chances of success, EVT makes assumption on the sampling distribution.

The usual assumptions may be formulated according to three (equivalent) viewpoints

- ▹ convergence in distribution of rescaled and recentered sample maxima
- ▹ convergence in distribution of rescaled excess distributions
- ▹ (extended) regular variation of the tail quantile function

Heavy tails: $\gamma > 0$

In the heavy tails (Frechet) domain, characterizations of the domains of attraction can be simplified in the two equivalent forms:

- Tail function is regularly varying with index $-1/\gamma$:

$$\lim_{t\to\infty} \frac{\overline{F}(tx)}{\overline{F}(t)} = x^{-1/\gamma}$$

- Tail quantile function is regularly varying with index γ

$$\lim_{t\to\infty} \frac{U(tx)}{U(t)} = x^{\gamma}$$

  where $U(t) = \inf\{x : F(x) \geq 1 - 1/t\}$.

Definition ($F \in \text{MDA}(\gamma)$)

F is said to belong the max-domain of attraction of index γ

Convergences can be arbitrarily slow

The tail index estimation problem

Assuming $F \in \mathrm{MDA}(\gamma)$, and given $X_1, \ldots, X_n \sim_{i.i.d.} F$, a basic inferential problem consists of estimating γ from the data

The fact that $F \in \mathrm{MDA}(\gamma)$ only impacts the tail of F.
Belonging to a domain of attraction is an asymptotic property of F or U

Inference on the tail index has to be based on the tail of the empirical distribution or equivalently on the tail empirical quantile function

Order statistics

$X_{(1)} \geq X_{(2)} \geq \ldots \geq X_{(n)}$ is the non-increasing rearrangement of the sample.

A semi-parametric problem

Assuming that U is γ-regularly varying leads to

Karamata's representation

$$U(t) = c(t) t^{\gamma} \exp \int_1^t \frac{\eta(s)}{s} \mathrm{d}s \qquad \lim_t c(t) = c \qquad \lim_s \eta(s) = 0$$

Von Mises conditions: $c(t) = c$

As the tail quantile function U characterizes the sampling distribution, this representation stresses the fact that tail index estimation fits into the framework of *semi-parametric inference* where

- the (finite-dimensional) tail index γ is the parameter of interest
- the (infinite dimensional) parameter (c, η) is the nuisance parameter

Peaks over thresholds

Tail index estimators are

Functions of largest order statistics

$$\widehat{\gamma}(k) = F(X_{(1)}, \ldots, X_{(k+1)})$$

Popular examples:

Hill estimator

$$\widehat{\gamma}(k) = \frac{1}{k}\sum_{i=1}^{k} \ln \frac{X_{(i)}}{X_{(k+1)}} = \frac{1}{k}\sum_{i=1}^{k} i \ln \frac{X_{(i)}}{X_{(i+1)}}$$

Works for $\gamma > 0$ (Fréchet domain)

Pickands estimator

$$\widehat{\gamma}(k) = \log_2 \frac{X_{(k)} - X_{(2k)}}{X_{(2k)} - X_{(4k)}}$$

Tail index estimators (cont'd)

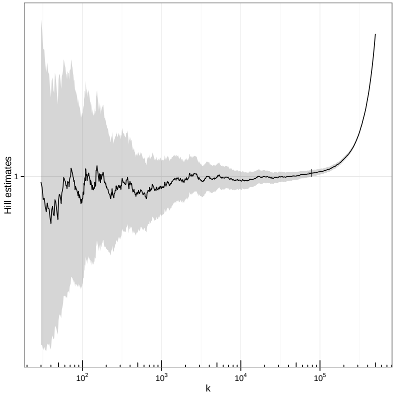
Each tail index estimator (Hill, Pickands, Moment, MLE, ...) is actually a sequence of estimators indexed by the number of order statistics used to compute the estimator.

Each sequence of estimators raises an *estimator selection* problem

EVT has developed visualization tools to help practitioners in picking reasonable number of order statistics.

The Hill plot and the *alternative* Hill plots are the canonical examples of such visualization tools

Alternative Hill plot



- Hill estimators for a Cauchy sample ($n = 10^6$)
- Extreme value index is 1
- Bias $\mathbb{E}b\left(e^{Y_{(k+1)}}\right)$ where

$$b(t) = t\int_t^\infty \frac{\eta(s)}{s^2}\mathrm{d}s$$

$$b(t) = O(1/t)$$

- The gray area is an approximate 95% confidence region $(\widehat{\gamma}(k)(1 + \pm z_{.95}/\sqrt{k}))$

EVT $\stackrel{?}{=}$ Asymptopia

Extreme value analysis

Learning theory has been a playground and a driving force for the development of concentration inequalities.
For a long time, Extreme Value Analysis (EVT) remained immune to the concentration of measure phenomenon

EVT has traditionally been regarded as an asymptotic theory

but the simplest (and most basic) statistical problems in EVT raise model selection issues that somehow could benefit from the availability of non-asymptotic tail bounds

Tail index estimators as smooth tail functionals

Asymptotic analysis establishes that under mild assumptions, tail index estimators recentered around their expectation and suitably rescaled are asymptotically Gaussian

Using the *exponential representation trick*, $\rightsquigarrow$ a non-asymptotic view at this smoothness property

This provides a (relatively) transparent analysis of a data-driven (would-be adaptive) strategy for choosing the number of order statistics used in tail index estimation

Risk profile comparison

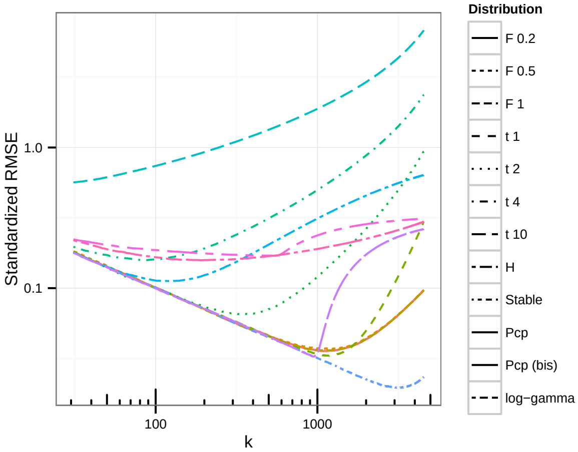


Figure: Monte-Carlo estimates of the standardised root mean square error (RMSE) of Hill estimators as a function of the number of order statistics k for samples of size 10000 from different sampling distributions.

Bias-variance tradeoff

Risk of Hill estimator

$$\mathbb{E}\left[(\gamma - \widehat{\gamma}(k))^2\right] = \mathrm{var}(\widehat{\gamma}(k)) + (\gamma - \mathbb{E}\widehat{\gamma}(k))^2$$

Best choice of k depends

on unknown $b(t) = \mathbb{E}[\gamma - \hat{\gamma}(k) \mid 1/\overline{F}(X_{(k+1)}) = t] = t \int_t^\infty \eta(s)/s^2 \mathrm{d}s$

Since 1980,

Many attempts to select asymptotically best possible k under a second order restriction (assuming the bias b is regularly varying with index $\rho < 0$).
Starts with Hall-Welsch (1985), Danielson et al. (1998), Drees and Kaufmann (1998), and too many to name ...

Recently

Attempts to derive risk bounds only assuming that b decays sufficiently fast (without assuming second order regularity).
Grama and Spokoiny (2008), Carpentier and Kim (2014)

A data-driven choice

Lepski's method

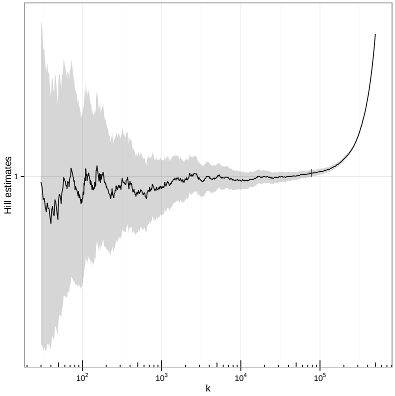
$$\widehat{k}(r_n) = \min\left\{k \in \{2,\dots,n\} : \max_{2\le i\le k} \sqrt{i}\frac{|\widehat{\gamma}(i) - \widehat{\gamma}(k)|}{\widehat{\gamma}(i)} > r_n(\delta)\right\}$$

with $\sqrt{\ln\ln n} = O(r_n)$ and $r_n = O(\sqrt{n})$

Suggestions for tuning $r_n(\delta)$

- $r_n(\delta) = 1$, and the ability to replace $\widehat{\gamma}(i)$ by $\mathbb{E}\widehat{\gamma}(i)$ would balance bias and standard deviation.
- $r_n(\delta) \ggg \sqrt{\ln\ln n}$ is too conservative
- $r_n(\delta) \propto \sqrt{\ln\ln n}$?

Alternative Hill plot



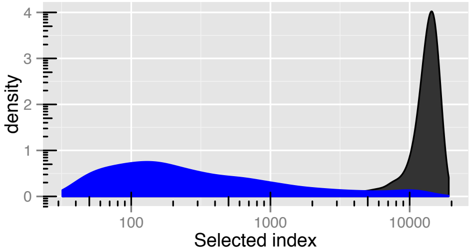
- Hill estimators for a Cauchy sample ($n = 10^6$)
- Extreme value index is 1
- Bias $\mathbb{E}b\left(e^{Y_{(k+1)}}\right)$ where

$$b(t) = t\int_t^\infty \frac{\eta(s)}{s^2}\mathrm{d}s$$

$$b(t) = O(1/t)$$

- The gray area is an approximate 95% confidence region $(\widehat{\gamma}(k)(1 + \pm z_{.95}/\sqrt{k}))$
- Point $+$ represents the index selected by Lepski's rule with $r_n = \sqrt{2\ln\ln n}$.

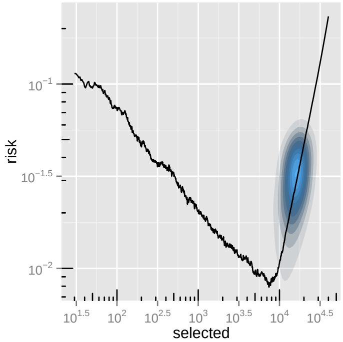
A calibration problem: Hill estimators on samples from Cauchy distribution ($\gamma = 1$).

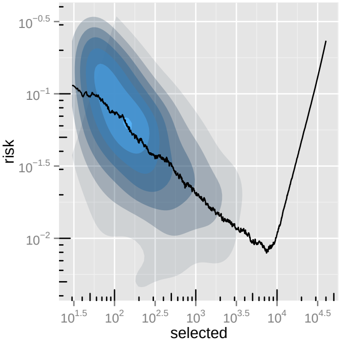


Distribution of selected estimates when the threshold is chosen as

- $\sqrt{3 \log \log(n)}$ (black area) favoring variance over bias
- $\sqrt{\log(\log(n))}$ (blue area) favoring bias over variance.

A calibration problem: Hill estimators on samples from Cauchy distribution ($\gamma = 1$).





The risk of each Hill estimator is estimated from 500 simulations.
Each simulation is performed on a sample of size 10^5.

The selection procedure is carried out with $r_n = \begin{cases} \sqrt{3 \ln \ln n} & \text{(left)} \\ \sqrt{\ln \ln n} & \text{(right)} \end{cases}$

A typical situation

Selecting a reasonable index k amounts to betting that the target distribution fits into a model defined by an envelope condition on the bias b ($\sqrt{k}b(n/k) \approx \gamma$ or rather $\sqrt{k}b(n/k) \approx r_n\gamma$) .

In order to provide risk bounds for this selection procedure, we need to be able to control the sequence of random variables $(\widehat{\gamma}(k) - \gamma)_k$ *over an appropriate range of* k*'s*.

Good concentration inequalities can be piped into union bounds.

Oracle inequalities: a generic result

A pivotal index: $k_n = \max\left\{k : \ell_n \leq k \leq n \text{ and } \sqrt{k}\,\overline{\eta}(n/k^\delta) \geq \gamma r_n\right\}$

with $r_n = \sqrt{c_2 \ln \ln n}$ and $k^\delta = k + \sqrt{2k \ln(1/\delta)} + 2\ln(1/\delta)$.

Oracle inequalities: a generic result

A pivotal index: $k_n = \max\left\{k : \ell_n \leq k \leq n \text{ and } \sqrt{k}\,\overline{\eta}(n/k^\delta) \geq \gamma r_n\right\}$

with $r_n = \sqrt{c_2 \ln\ln n}$ and $k^\delta = k + \sqrt{2k\ln(1/\delta)} + 2\ln(1/\delta)$.

A threshold: $r_n(\delta) = 10 r_n + (1 + 3r_n/\sqrt{\ell_n})\left(\xi_n + \sqrt{8\ln(2/\delta)} + \frac{\ln(2/\delta)}{\sqrt{\ell_n}}\right)$

with $\xi_n = c_1\sqrt{\ln\log_2(n)} + c_1'$ and $\widehat{k}_n = \widehat{k}(r_n(\delta))$

Oracle inequalities: a generic result

A pivotal index: $k_n = \max\left\{k : \ell_n \leq k \leq n \text{ and } \sqrt{k}\,\overline{\eta}(n/k^\delta) \geq \gamma r_n\right\}$

with $r_n = \sqrt{c_2 \ln\ln n}$ and $k^\delta = k + \sqrt{2k\ln(1/\delta)} + 2\ln(1/\delta)$.

A threshold: $r_n(\delta) = 10 r_n + (1 + 3r_n/\sqrt{\ell_n})\left(\xi_n + \sqrt{8\ln(2/\delta)} + \frac{\ln(2/\delta)}{\sqrt{\ell_n}}\right)$

with $\xi_n = c_1\sqrt{\ln\log_2(n)} + c_1'$ and $\widehat{k}_n = \widehat{k}(r_n(\delta))$

Assume that n is large enough so that k_n is well-defined

For $2/n < \delta < 1/4$, with probability larger than $1 - 3\delta$,

$$\left|\gamma - \widehat{\gamma}(\widehat{k}_n)\right| \leq \left|\gamma - \widehat{\gamma}(k_n)\right|\left(1 + \frac{r_n(\delta)}{\sqrt{k_n}}\right) + \frac{r_n(\delta)\gamma}{\sqrt{k_n}} .$$

And, with probability larger than $1 - 4\delta$,

$$\left|\widehat{\gamma}(\widehat{k}_n) - \gamma\right| \leq \frac{2r_n(\delta)}{\sqrt{k_n}}\gamma\left(1 + \alpha(\delta, n)\right) \text{ with } \alpha(\delta, n) \leq \frac{r_n}{2\sqrt{\ell_n}} + \frac{\sqrt{\ln(2/\delta)}}{r_n(\delta)}\left(1 + \frac{3r_n(\delta)}{\sqrt{\ell_n}}\right)^2$$

Lower bounds: adaptivity has a price (Carpentier & Kim (2014))

Let $\rho < -1$. Let v belong to $(0, \mathrm{e}/((1+2\,\mathrm{e}))$. For any tail index estimator $\widehat{\gamma}$, for any sample size n such that $\lfloor \ln n \rfloor > \mathrm{e}/v$, let $M = \lfloor \ln n \rfloor$, then there exists a probability distribution $P \in \text{MDA}(\gamma), \gamma > 0$ satisfying the von Mises condition with von Mises function η satisfying

$$\overline{\eta}(t) \leq t^{\rho}$$

where $\rho_0 \leq \rho < 0$ and such that

$$P^{\otimes n}\left\{|\widehat{\gamma} - \gamma| \geq \frac{\kappa_\rho}{4}\gamma\left(\frac{v \ln \ln n}{n}\right)^{|\rho|/(1+2|\rho|)}\right\} \geq \frac{1}{1+2\,\mathrm{e}}$$

and

$$\mathbb{E}_P\left[\frac{|\widehat{\gamma} - \gamma|}{\gamma}\right] \geq \frac{\kappa_\rho}{4(1+2\,\mathrm{e})}\left(\frac{v \ln \ln n}{n}\right)^{|\rho|/(1+2|\rho|)},$$

with $\kappa_\rho = \exp(-1/(1+2|\rho|)^2)$.

If ρ were known (insider deal), we could get rid of $(\ln \ln n)^{|\rho|/(1+2|\rho|)}$.

Oracle inequalities: Enveloppe conditions for $\overline{\eta}$

There exists a constant $\kappa_{C,\delta,\rho}$ depending on $C > 0, \delta > 0$ and $\rho < 0$ such that:

If for some $C > 0$ and $\rho < 0$,

$$\overline{\eta}(t) \leq Ct^{\rho} ,$$

then, with probability larger than $1 - 4\delta$,

$$\left|\widehat{\gamma}(\widehat{k}_n) - \gamma\right| \quad \leq \quad \kappa_{C,\delta,\rho}\left(\frac{\gamma^2 \ln\left((2/\delta)\ln n\right)}{n}\right)^{|\rho|/(1+2|\rho|)} (1 + \alpha(\delta, n))$$

Exponential representation tricks

Exponential representation

Exponential representation of order statistics

Quantile transform

$U(t) := \overline{F}(1-1/t)$ for $t > 1$

$$X \sim F \quad \Rightarrow \quad X \sim U(\exp(Y)) : \qquad \text{with } Y \sim \text{Exponential}$$

Karamata's representation

$$U(t) = c(t) t^\gamma \exp \int_1^t \frac{\eta(s)}{s} \mathrm{d}s \qquad \lim_t c(t) = c \qquad \lim_s \eta(s) = 0$$

Von Mises conditions: $c(t) = c$

Rényi's representation

Order statistics of exponential samples are partial sums

$$(Y_{(1)} \geq \ldots \geq Y_{(n)}) \sim \left(\sum_{i=k}^n \frac{E_i}{i} \right)_{k \in \{1,\ldots,n\}} \qquad \text{where} \quad E_i \sim_{\text{i.i.d.}} \text{Exponential}$$

Exponential representation of Order Statistics (cont'd)

Consequences

$$\left(X_{(1)}, X_{(2)}, \ldots, X_{(n)}\right) \sim \left(U\left(\mathrm{e}^{Y_{(1)}}\right), U\left(\mathrm{e}^{Y_{(2)}}\right), \ldots, U\left(\mathrm{e}^{Y_{(n)}}\right)\right)$$

where $Y_{(1)}, \ldots, Y_{(n)}$ are order statistics of an exponential sample.

$$\left(X_{(1)}, X_{(2)}, \ldots, X_{(n)}\right) \sim \left(U\left(\mathrm{e}^{\sum_{i=1}^{n} E_i/i}\right), U\left(\mathrm{e}^{\sum_{i=2}^{n} E_i/i}\right), \ldots, U\left(\mathrm{e}^{E_n/n}\right)\right)$$

where E_i are i.i.d. exponential

$$X_{(i)} \sim c\,\mathrm{e}^{\gamma \sum_{j=i}^{n} E_j/j} \exp\left(\int_0^{\sum_{j=i}^{n} E_j/j} \eta\left(\mathrm{e}^s\right) \mathrm{d}s\right)$$

$$\ln X_{(i)} = \ln c + \gamma \sum_{j=i}^{n} \frac{E_j}{j} + \int_0^{\sum_{j=i}^{n} E_j/j} \eta\left(\mathrm{e}^s\right) \mathrm{d}s$$

Hill estimators as functions of independent exponential random variables

Combining Rényi's representation, Karamata's representation, the quantile transform and the Von Mises condition ...

Hill estimators as functionals of an exponential sample

$$\left(\widehat{\gamma}(k)\right)_{2\leq k\leq n} \sim \left(\frac{1}{k}\sum_{i=1}^{k}\int_0^{E_i}\left(\gamma+\eta(\mathrm{e}^{\frac{u}{i}+Y_{(i+1)}})\right)\mathrm{d}u\right)_{k<n}$$

where $E_1,\ldots,E_n \sim_{i..i.d.}$ standard exponentials, and $Y_{(k)} = \sum_{j=k}^{n} E_j/j$

Hill estimators are *approximately* distributed like partial sums of independent exponential random variables

Overlooked concentration inequalities

Talagrand's inequality (early 90's)

Concentration for smooth functional of exponential samples

If f is a differentiable function on $\mathbb{R}^n$ with $\max_i |\partial_i f| < \infty$, and $Z = f(E_1, \ldots, E_n)$ where $E_1, \ldots, E_n$ are independent standard exponential random variables.

- Let $c < 1$, then for all λ such that $0 \leq \lambda \max_i |\partial_i f| \leq c$,

$$\mathrm{Ent}\left[e^{\lambda(Z-\mathbb{E}Z)}\right] \leq \frac{2\lambda^2}{1-c}\mathbb{E}\left[e^{\lambda(Z-\mathbb{E}Z)}\,\|\nabla f\|^2\right]$$

- Let v be the essential supremum of $\|\nabla f\|^2$. Then,
  Z is sub-gamma on both tails with variance factor $4v$ and scale factor $\max_i |\partial_i f|$

Maurey (1992) described an alternative proof using infimum -convolution arguments. Bobkov and Ledoux (1997) gave a proof based on the entropy method. The result generalizes to log-concave distributions

Concentration/deviation inequalities for Hill process

$T := \exp(Y_{(J+1)})$
For some k, $c_3 \ln n \vee 32 \leq \ell \leq k \leq J$ where $c_3 \geq 2$,

$$Z^a := \max_{\ell \leq i \leq k} \sqrt{i} \left| \widehat{\gamma}(i) - \mathbb{E}\left[\widehat{\gamma}(i) \mid Y_{(k+1)}\right] \right| .$$

Conditionally on T,

i) For $\ell \leq i \leq k$,

$$\sqrt{i}\left(\widehat{\gamma}(i) - \mathbb{E}[\widehat{\gamma}(i) \mid T]\right) \in \Gamma_{\pm}\left(4\left(\gamma + 3\overline{\eta}(T)\right)^2, \left(\gamma + 3\overline{\eta}(T)\right)\right) .$$

ii) Assume that $J\overline{\eta}(\exp(Y_{(J+1)}))^2 \leq \gamma^2 r_n^2$ where $r_n > \sqrt{c_2 \ln \ln n}$ with $c_2 = 2$. Then

$$Z^a \in \Gamma_{\pm}\left(4\gamma^2(1 + 3r_n/\sqrt{J})^2, \gamma(1 + 3r_n/\sqrt{J})/\sqrt{\ell}\right)$$

and

$$\mathbb{E}\left[Z^a \mid T\right] \leq \gamma \xi_n \left(1 + \frac{3r_n}{\sqrt{J}}\right) .$$

The End

References

- J. Beirlant, Y. Goegebeur, J. Teugels, and J. Segers. *Statistics of extremes.* John Wiley & Sons, Ltd., 2004.
- S. Boucheron and M. Thomas. Tail index estimation, concentration and adaptivity. http://arxiv.org/abs/1503.05077
- A. Carpentier and A. Kim. Adaptive and minimax optimal estimation of the tail coefficient. http://arxiv.org/pdf/1309.2585.pdf, 2014a.
- H. Drees and E. Kaufmann. Selecting the optimal sample fraction in univariate extreme value estimation. Stoch. Process and Appl. 1998.
- I. Grama and V. Spokoiny. Statistics of Extremes by Oracle Estimation. AoS 2008.
- L. de Haan & A. Ferreira. *Extreme value theory.* Springer, 2006.
- M. Ledoux. *The concentration of measure phenomenon*. AMS, 2001.
- S. Resnick. *Heavy tails*. Springer, 2007.