Analysis methods of heavy-tailed data

Natalia Markovich

Institute of Control Sciences
Russian Academy of Sciences, Moscow, Russia

February, 13-18, 2006, Bamberg, Germany
June, 19-23, 2006, Brest, France
May, 14-19, 2007, Trondheim, Norway
PhD course
Introduction:
definitions and basic properties of classes of heavy-tailed distributions. Tail index estimation. Methods for the selection of the number of the largest order statistics in Hill estimator. Rough methods for the detection of heavy tails and the number of finite moments.

The Section 1 contains the introduction
with necessary definitions, basic properties and examples of heavy-tailed data. The tail index indicates the shape of the tail and therefore it is the basic characteristic of heavy-tailed data. Methods of tail index estimation are presented. Finally, several rough tools for the detection of heavy-tails, the number of finite moments and dependence are considered.
Main assumption.

Let $X_1, \ldots, X_n$ be a sample of i.i.d. r.v.s distributed with the heavy-tailed $DF F(x)$.

**Definition**

A $DF F(x)$ (or the r.v. $X$) is called heavy-tailed if its tail $\bar{F}(x) = 1 - F(x) > 0, x \geq 0,$ satisfies $\forall y \geq 0$

$$\lim_{x \to \infty} P\{X > x + y | X > x\} = \lim_{x \to \infty} \frac{\bar{F}(x + y)}{\bar{F}(x)} = 1.$$
### Examples of heavy- and light-tailed distributions

<table>
<thead>
<tr>
<th>Heavy-tailed distributions</th>
<th><strong>Subexponential:</strong> Pareto, Lognormal, Weibull with shape parameter less than 1. <strong>With regularly varying tails:</strong> Pareto, Cauchy, Burr, Frechét, Zipf-Mandelbrot law.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Light-tailed distributions</td>
<td>exponential, gamma, Weibull with shape parameter more than 1, normal, finite distributions.</td>
</tr>
</tbody>
</table>
Heavy-tailed distributions have been accepted as realistic models for various phenomena:

- WWW-session characteristics
  - sizes and durations of sub-sessions; sizes of responses
  - inter-response time intervals
- on/off-periods of packet traffic
- file sizes
- service-time in queueing model
- flood levels of rivers
- major insurance claims
- extreme levels of ozon concentrations
- high wind-speed values
- wave heights during a storm
- low and high temperatures
Basic definitions

Let \( X^n = \{X_1, \ldots, X_n\} \) be a sample of i.i.d. r.v. distributed with the \( DF \, F(x) \) and \( M_n = \max(X_1, X_2, \ldots, X_n) \).

Extreme value theory assumes that for a suitable choice of normalizing constants \( a_n, b_n \) it holds

\[
P\left\{ \frac{(M_n - b_n)}{a_n} \leq x \right\} = F^n(b_n + a_n x) \to_{n \to \infty} H_\gamma(x), \, x \in R,
\]

and an Extreme Value DF \( H_\gamma(x) \) is of the following type:

\[
H_\gamma(x) = \begin{cases} 
\exp(-x^{-1/\gamma}), & x > 0, \gamma > 0 \quad \text{’Fréchet’,} \\
\exp((-x)^{-1/\gamma}), & x < 0, \gamma < 0 \quad \text{’Weibull’,} \\
\exp(-e^{-x}), & \gamma = 0, x \in R \quad \text{’Gumbel’}. 
\end{cases}
\]

Definition

The parameter \( \gamma \) is called the extreme value index (EVI) and defines the shape of the tail of the r.v. \( X \).

The parameter \( \alpha = 1/\gamma \) is called tail index.
Pickands’s theorem:

The limit distribution of the excess distribution of the $X_i$’s is necessarily of the Generalized Pareto form

$$
\lim_{u \uparrow x_F, u+x < x_F} \mathbb{P}(X_1 - u > x | X_1 > u) \to (1 + \gamma x)_+^{-1/\gamma}, \quad x \in \mathbb{R},
$$

where

$$
x_F = \sup\{x \in \mathbb{R} : F(x) < 1\}
$$

is the right endpoint of the distribution $F$ and the shape parameter $\gamma \in \mathbb{R}$.

Therefore, the Generalized Pareto distribution (GPD) with DF

$$
\Psi_{\sigma, \gamma}(x) = 1 - (1 + \gamma x / \sigma)_+^{-1/\gamma}
$$

is often used as a model of the tail of the distribution.
The classes of heavy-tailed distributions

- **distributions with regularly varying tails (RVT)**

  \( X \in \mathcal{R}_{-1/\gamma} \)

  \[ P\{X > x\} = x^{-1/\gamma} \ell(x), \forall x > 0, \gamma > 0, \]

  where \( \ell \) is a slowly varying function, i.e.

  \[ \lim_{x \to \infty} \frac{\ell(tx)}{\ell(x)} = 1, \quad \forall t > 0. \]

**Examples:**

Pareto, Burr, Fréchet distributions belong to RVT.

**Examples of \( \ell(x) \) give**

\( c \ln x, c \ln(\ln x) \) and all functions converging to positive constants.
The classes of heavy-tailed distributions

- **subexponential distributions** ($S$)
  
  \[ (X \in S) \]

  \[ P\{S_n > x\} \sim nP\{X_1 > x\} \sim P\{M_n > x\} \quad \text{as} \quad x \to \infty, \]

  where \( S_n = X_1 + \ldots + X_n, \ n \geq 2, \ M_n = \max_{i=1,\ldots,n}\{X_i\}. \]

**Example:**

Weibull with shape parameter less than 1 belongs to $S$.

**Intuitively, subexponentiality means**

that the only way the sum can be large is by one of the summands getting large (in contrast, in the light-tailed case all summands are large if the sum is so).
Basic properties of regularly varying distributions:

**Lemma**

Let \( X \in \mathcal{R}_{-\alpha} \). Then,

(i) \( X \in S \).

(ii) \( \mathbb{E}\{X^\beta\} < \infty \) if \( \beta < \alpha \), \( \mathbb{E}\{X^\beta\} = \infty \) if \( \beta \geq \alpha \).

(iii) If \( \alpha > 1 \), then \( X^r \in \mathcal{R}_{1-\alpha} \) and

\[
\mathbb{P}\{X^r > x\} \sim \ell(x)x^{1-\alpha}/((\alpha - 1)\mathbb{E}\{X\}) \quad \text{as} \quad x \to \infty.
\]

(iv) If \( Y \) is non-negative and independent of \( X \) such that

\[
\mathbb{P}\{Y > x\} = \ell_2(x)x^{-\alpha_2}, \quad \text{then } X + Y \in \mathcal{R}_{-\min(\alpha, \alpha_2)} \text{ and }
\]

\[
\mathbb{P}\{X + Y > x\} \sim \mathbb{P}\{X > x\} + \mathbb{P}\{Y > x\} \quad \text{as } x \to \infty.
\]

(v) If \( Y \) is non-negative and independent of \( X \) such that

\[
\mathbb{E}\{Y^{\alpha + \varepsilon}\} < \infty \text{ for some } \varepsilon > 0 \text{ then } XY \in \mathcal{R}_{-\alpha} \text{ and }
\]

\[
\mathbb{P}\{XY > x\} \sim \mathbb{E}\{Y^{\alpha}\}\mathbb{P}\{X > x\} \quad \text{as} \quad x \to \infty.
\]
Important property for the rough detection of heavy tails and the number of finite moments:

Let $X \in \mathcal{R}_{-\alpha}$.

Then $\mathbb{E}\{X^\beta\} < \infty$, if $\beta < 1/\gamma$;  \hspace{1cm} \mathbb{E}\{X^\beta\} = \infty$, if $\beta > 1/\gamma$.

**Examples:**

- If $\alpha = 2$, $\gamma = 0.5$, then
  \[ \mathbb{E}X_1 < \infty \text{ (the first moment is finite)}, \]
  \[ \mathbb{E}X_1^2 = \infty \text{ (the second moment is infinite or does not exist)}. \]

- If $\alpha = 0.5$, $\gamma = 2$, then
  \[ \mathbb{E}X_1 = \infty, \mathbb{E}X_1^2 = \infty \ldots \text{ (all moments are infinite or do not exist)}. \]
Estimators of tail index

\[ X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)} \]

are order statistics of the sample \( X^n = \{X_1, X_2, \ldots, X_n\} \)

For \( \gamma > 0 \):

- **Hill’s estimator**

\[
\hat{\gamma}^H(n, k) = \frac{1}{k} \sum_{i=1}^{k} \ln X_{(n-i+1)} - \ln X_{(n-k)}
\]


\[
a_n = a_n(x_n) = \frac{\sum_{i=1}^{n} \ln(X_i/x_n)I\{X_i > x_n\}}{\sum_{i=1}^{n} I\{X_i > x_n\}},
\]

\( x_n \) is an arbitrary threshold level, e.g., \( x_n = X_{(n-k)} \)
### Estimators of tail index

For $\gamma \in R$:

  
  $\hat{\gamma}^M(n, k) = \hat{\gamma}^H(n, k) + 1 - 0.5 \left( 1 - (\hat{\gamma}^H(n, k))^2 / S_{n,k} \right)^{-1}$,

  $S_{n,k} = (1/k) \sum_{i=1}^{k} \left( \ln X_{(n-i+1)} - \ln X_{(n-k)} \right)^2$

  
  $\hat{\gamma}^{UH}(n, k) = (1/k) \sum_{i=1}^{k} \ln UH_i - \ln UH_{k+1}$,

  $UH_i = X_{(n-i)} \hat{\gamma}^H(n, i)$

- **Pickands's estimator**
  
  $\hat{\gamma}^P(n, k) = 1 / \ln 2 \ln \left( X_{(k)} - X_{(2k)} \right) / \left( X_{(2k)} - X_{(4k)} \right)$

No recursiveness!
The sample $X^n$ is divided into $l$ groups $V_1, \ldots, V_l$, each group containing $m$ r.v.s, i.e. $n = l \cdot m$.

**Estimator of the function of the tail index:**

$$z_l = \frac{1}{l} \sum_{i=1}^{l} k_{li} = \frac{\hat{\alpha}}{\hat{\alpha}+1} = \frac{1}{1+\hat{\gamma}},$$

where

$$k_{li} = \frac{M_{li}^{(2)}}{M_{li}^{(1)}}, \quad M_{li}^{(1)} = \max\{X_j : X_j \in V_i\}$$

and $M_{li}^{(2)}$ is the second largest element in the same group $V_i$. 
For distributions with regularly varying tails

\[ 1 - F(x) = x^{-\alpha} \ell(x), \]

and \( l = m = \lfloor \sqrt{n} \rfloor \), it is proved

\[ z_l \rightarrow^{a.s.} \frac{\alpha}{\alpha + 1} = \frac{1}{1 + \gamma}. \]

For distributions

\[ 1 - F(x) = C_1 x^{-\alpha} + C_2 x^{-\beta} + o(x^{-\beta}), \]

with \( \beta = 2\alpha \) it holds

\[ l(l^{-1} \sum_{i=1}^l k_{li} - \alpha(1+\alpha)^{-1}) \left( \sum_{i=1}^l k_{li}^2 - l^{-1} \sum_{i=1}^l k_{li} \right)^{-1/2} \rightarrow^p N(0, 1). \]
Recursiveness of the estimate $z_l$. On-line estimation.

Having the next group of observations $V_{l+1}$ it follows

$$\gamma_{l+1} = \left( \frac{l}{l+1} \cdot \frac{1}{1 + \gamma_l} + \frac{k_{l+1,l+1}}{l+1} \right)^{-1} - 1$$

After getting $i$ additional groups with $m$ elements each $V_{l+1}, \ldots, V_{l+i}$

$$\gamma_{l+i} = (l + i) \left( \frac{l}{1 + \gamma_l} + k_{l+1,l+1} + \ldots + k_{l+i,l+i} \right)^{-1} - 1,$$

i.e. $\gamma_{l+i}$ is obtained using $\gamma_l$ by $O(1)$ calculations.
Recursiveness of the estimate $z_l$. On-line estimation.

Since

$$z_{l+i} = \frac{1}{1 + \gamma_{l+i}}$$

it holds

$$z_{l+i} = \left( l z_l + \sum_{j=1}^{i} k_{l+j+l} \right) / (l + i),$$

$$bias(z_{l+i}) = bias(z_l), \quad var(z_{l+i}) = var(z_l)l / (l + i),$$

$$var(z_{l+i}) < var(z_l) \quad \text{for} \quad \forall i > 0$$
The selection of the number of the largest order statistics in Hill estimator

- **A Hill plot** \{\((k, \hat{\gamma}^H(n, k)), 1 \leq k \leq n - 1\)\}: the estimate of \(\hat{\gamma}^H(n, k)\) is chosen from an interval in which this function demonstrate stability.

- **Plot of the mean excess function** \{(\(u, e(u)\) : \(X_1 < u < X_n\)\}, where

\[
e(u) = \frac{\sum_{i=1}^{n} (X_i - u)1\{X_i > u\}}{\sum_{i=1}^{n} 1\{X_i > u\}}
\]

is the sample mean excess function over the threshold \(u\).
If this plot follows a reasonably straight line above a certain value of \(u\), then this indicates that excesses over \(u\) follow
\[
e^P(u) = \frac{1 + \gamma u}{1 - \gamma}
\]

of generalized Pareto distribution with positive shape parameter. One can take the nearest order statistics to \(u\) as the estimate of \(k\).
The Hill's estimate against $k$ for 15 samples of Weibull (left), Pareto (middle) and Frechét (right) distributions, all with shape parameter $\alpha = 1/\gamma = 0.5$. Sample size is $n = 1000$. 

The sensitivity of the Hill estimate to $k$. 

Natalia Markovich  
Analysis methods of heavy-tailed data
Plot of the mean excess function.

Left: Weibull distribution. Right: Pareto distribution. The shape parameter is 0.5, sample size 1000.
Bootstrap method for \( k \) selection.

- Minimizing of the empirical bootstrap estimate of the mean squared error of \( \gamma \):

\[
MSE(\hat{\gamma}) = E (\hat{\gamma} - \gamma)^2 \rightarrow \min_k.
\]

- The bootstrap estimate is obtained by drawing \( B \) samples with replacement from the original data set \( X^n \).

- Smaller re-samples \( \{X^*_1, ..., X^*_n\} \) of the size

\[
n_1 = n^\beta, \quad 0 < \beta < 1,
\]

are used.

- The corresponding smaller \( k_1 \) and an optimal \( k \) are related by:

\[
k = k_1 (n/n_1)^\alpha, \quad 0 < \alpha < 1,
\]

where \( \beta = 1/2 \) and \( \alpha = 2/3 \), P.Hall, (1990).
Empirical bootstrap estimate of the $MSE(\hat{\gamma})$ is

\[ MSE^*(n_1, k_1) = \left( \hat{b}^*(n_1, k_1) \right)^2 + \hat{\text{var}}^*(n_1, k_1) \to \min_{k_1}, \]

where

\[ \hat{b}^*(n_1, k_1) = \frac{1}{B} \sum_{b=1}^{B} \hat{\gamma}^*_b(n_1, k_1) - \hat{\gamma}(n, k), \]

\[ \hat{\text{var}}^*(n_1, k_1) = \frac{1}{B-1} \sum_{b=1}^{B} \left( \hat{\gamma}^*_b(n_1, k_1) - \frac{1}{B} \sum_{b=1}^{B} \hat{\gamma}^*_b(n_1, k_1) \right)^2 \]

are the empirical bootstrap estimates of the bias and the variance,

$\hat{\gamma}^*_b$ is the Hill’s estimate constructed by some re-sample of the size $n_1$ with the parameter $k_1$. 
Danielsson, de Haan, Peng and de Vries, (1997) requires less parameters than bootstrap method, Hall, (1990): $n_1$ and $B$ are required, $\alpha$ is not required.

**Auxiliary statistic:**

$$MSE(z_{n,k}) = E \left( z_{n,k} - z_{n,k}^* \right)^2,$$

where

$$z_{n,k} = M_{n,k} - 2(\hat{\gamma}^H(n, k))^2,$$

$$M_{n,k} = \frac{1}{k} \sum_{j=1}^{k} (\log X_{(n-j+1)} - \log X_{(n-k)})^2,$$

$z_{n,k}^*$ is a bootstrap estimate of $z_{n,k}$. 
Double bootstrap method for $k$ selection.

$M_{n,k}/2\hat{\gamma}^H(n,k)$ and $\hat{\gamma}^H(n,k)$ are consistent estimates for $\gamma$, 

$\Rightarrow z_{n,k} \to 0 \quad \text{as} \quad n \to \infty.$

$\Rightarrow AMSE(z_{n,k}) = \mathbf{E}(z_{n,k})^2 \to \min_k.$
Double bootstrap procedure is

- **Draw** $B$ bootstrap sub-samples of the size $n_1 \in (\sqrt{n}, n)$ (e.g., $n_1 \sim n^{3/4}$) from the original sample and determine the value $\hat{k}_{n_1}^*$ that minimizes $MSE$ of $z_{n_1,k}$.

- **Repeat** this for $B$ sub-samples of the size $n_2 = [n_1^2/n]$ ([x] is the integer part of the number) and determine the value $\hat{k}_{n_2}^*$ that minimizes $MSE$ of $z_{n_2,k}$.

- **Calculate** $\hat{k}_n^{opt}$ from the formula

\[
\hat{k}_n^{opt} = \left[ \left( \frac{\hat{k}_{n_1}^*}{\hat{k}_{n_2}^*} \right)^2 \left( 1 - \frac{1}{\hat{\rho}_1} \right)^{2\hat{\rho}_1^{-1}} \right], \quad \hat{\rho}_1 = \frac{\log \hat{k}_{n_1}^*}{2 \log(\hat{k}_{n_1}^*/n_1)},
\]

and estimate $\gamma$ by the Hill’s estimate with $\hat{k}_n^{opt}$.

The method is **robust** with respect to the choice of $n_1$, Gomes, Oliveira, (2000).
Sequential procedure for $k$ selection

is based on the theoretical result:

$$\sqrt{i}(\hat{\gamma}^H(n, i) - \gamma) \sim (\log \log n)^{1/2}, \quad 2 \leq i \leq k$$

in probability, **Drees & Kaufmann, (1998).**
Sequential procedure for $k$ selection. **Algorithm.**

- An initial estimate $\hat{\gamma}_0 = \hat{\gamma}^H(n, 2\sqrt{n})$ for the parameter $\gamma$ is obtained by the Hill’s estimate.
- For $r_n = 2.5\hat{\gamma}_0 n^{0.25}$ we compute
  \[
  \hat{k}(r_n) = \min\{k \in 1, \ldots, n - 1 | \max_{2 \leq i \leq k} \sqrt{i(\hat{\gamma}^H(n, i) - \hat{\gamma}^H(n, k))} > r_n\}.
  \]
  If $r_n$ is too large and $\max_{2 \leq i \leq k} \sqrt{i(\hat{\gamma}^H(n, i) - \hat{\gamma}^H(n, k))} > r_n$ is not satisfied it is recommended repeatedly replace $r_n$ by $0.9r_n$ until $\hat{k}(r_n)$ is well defined.
- Similarly, $\hat{k}(r^\varepsilon_n)$ for $\varepsilon = 0.7$ is computed.
- Optimal $k$
  \[
  \hat{k}^{opt} = \frac{1}{3} \left( \frac{\hat{k}(r^\varepsilon_n)}{(\hat{k}(r_n))^\varepsilon} \right)^{1/(1-\varepsilon)} (2\hat{\gamma}_0)^{1/3}
  \]
  is calculated and $\gamma$ is estimated by $\hat{\gamma}^H(n, \hat{k}^{opt})$.

The method is sensitive to the choice of $r_n$. 

Natalia Markovich
Analysis methods of heavy-tailed data
Plot method for $m$ selection in the group estimator.

The plot $\{(m, 1/z_m - 1)\}$ for Pareto distribution with $\gamma = 1$, the true $\gamma$ is shown by dotted line. Sample sizes are $n = \{150, 500, 1000\}$. 
Plot method for $m$ selection in the group estimator.

Plot: \( \{(m, z_m), m_0 < m < M_0\}, m_0 > 2, M_0 < n/2 \)

\[ m = n/l, \quad z_m = (m/n) \sum_{i=1}^{[n/m]} k(n/m)_i \]

From consistency result \( z_l \rightarrow^{a.s.} \frac{1}{1+\gamma} \) it follows that there must be an interval \([m_-, m_+]\)

such that \( z_m \approx \alpha/(1 + \alpha) = (1 + \gamma)^{-1} \) for all \( m \in [m_-, m_+] \).

We suggest choosing the average value

\[ \bar{z} = \text{mean}\{1/z_m - 1 : m \in [m_-, m_+]\} \]

and \( m^* \in [m_-, m_+] \) as a point such that \( z_{m^*} = \bar{z} \).
Bootstrap method for \(m\) automatical selection.

- Minimizing of the empirical bootstrap estimate of the mean squared error of \((1 + \gamma_l)^{-1}, l = n/m:\)

\[
MSE(\gamma_l) = E \left( \frac{1}{l} \sum_{i=1}^{l} k_{li} - \frac{1}{1 + \gamma} \right)^2 \rightarrow \min_m.
\]

- The bootstrap estimate is obtained by drawing \(B\) samples with replacement from the original data set \(X^n\).

- Smaller re-samples \(\{X^*_1, \ldots, X^*_n\}\) of the size

\[
n_1 = n^d, \quad 0 < d < 1,
\]

are used.

- The re-sample is divided into \(l_1\) subgroups.

- The size of subgroups \(m_1\) and \(m\) are related by:

\[
m = m_1 (n/n_1)^c, \quad 0 < c < 1,
\]

where \(m_1\) is the size of subgroups in re-samples.
Empirical bootstrap estimate of the MSE

\[
MSE^*(l_1, m_1) = \left( \hat{b}^*(l_1, m_1) \right)^2 + \hat{\text{var}}^*(l_1, m_1) \to \min_{m_1},
\]

where

\[
\hat{b}^*(l_1, m_1) = \frac{1}{B} \sum_{b=1}^{B} z_{l_1}^b - z_l,
\]

\[
\hat{\text{var}}^*(l_1, m_1) = \frac{1}{B - 1} \sum_{b=1}^{B} \left( z_{l_1}^b - \frac{1}{B} \sum_{b=1}^{B} z_{l_1}^b \right)^2
\]

are the empirical bootstrap estimates of the bias and the variance,

\[z_{l_1}^b = \frac{1}{l_1} \sum_{i=1}^{l_1} k_{l_1,i}\]

is the group estimator constructed by some re-sample.

How to select \(c\) and \(d\)?
Simulation study: the selection of $c$ and $d$.

- Asymptotic theory (P. Hall, (1990)) recommends $d = 1/2$ and $c = 2/3$ for the bootstrap estimation of the parameter $k$ of the Hill’s estimate of $\gamma$.

- We check $c = \{0.05, 0.1(0.1); 0.5\}$ for a fixed $d = 0.5$.

Relative bias and the square root of the mean squared error:

$$
Bias_\gamma = \frac{1}{\gamma} \left( \frac{1}{N_R} \sum_{i=1}^{N_R} \hat{\gamma}_i - \gamma \right),
$$

$$
RMSE_\gamma = \frac{1}{\gamma} \sqrt{\frac{1}{N_R} \sum_{i=1}^{N_R} (\hat{\gamma}_i - \gamma)^2}
$$
Conclusions:

- the best values of $c$ for the fixed $d = 0.5$ are $c = \{0.3 \div 0.5\}$;
- the bias of the group estimator is larger for Weibull distribution.

Further research:

- proof of theoretically best values of $c$ and $d$ for the group estimator using the bootstrap.
Rough methods for the detection of heavy tails and the number of finite moments.

1. **Ratio of the maximum to the sum**
   Let $X_1, X_2, \ldots, X_n$ be i.i.d. r.v.s. We define the following statistic
   
   $$ R_n(p) = \frac{M_n(p)}{S_n(p)}, \quad n \geq 1, \quad p > 0, \quad where $$

   $$ S_n(p) = |X_1|^p + \ldots + |X_n|^p, $$
   $$ M_n(p) = \max (|X_1|^p, \ldots, |X_n|^p), \quad n \geq 1 $$

   to check the moment conditions of the data.

   For different values of $p$ the plot of $n \to R_n(p)$ gives a preliminary information about the distribution $P\{|X| > x\}$. Then $E|X|^p < \infty$ follows, if $R_n(p)$ is small for large $n$. For large $n$ a significant difference between $R_n(p)$ and zero indicates that the moment $E|X|^p$ is infinite.
2. **QQ-plot** A QQ-plot (or “quantiles against quantiles”-plot) draws the dependence
\[
\{ \left( X_{(k)}, F^{-1} \left( \frac{n-k+1}{n+1} \right) \right) : k = 1, \ldots, n \},
\]
where \( X_{(1)} \geq \ldots \geq X_{(n)} \) are the order statistics of the sample, and \( F^{-1} \) is an inverse function of the DF \( F \).

Often, the QQ-plot is built as a dependence of exponential quantiles against the order statistics of the underlying sample. Then \( F^{-1} \) is an inverse function of the exponential DF. Then a linear QQ-plot corresponds to the exponential distribution.
3. Plot of the mean excess function

\[ e_n(u) = \frac{\sum_{i=1}^{n} (X_i - u) \mathbb{1}\{X_i > u\}}{\sum_{i=1}^{n} \mathbb{1}\{X_i > u\}} \]

1. Heavy-tailed distributions: \( e(u) \rightarrow \infty, \ u \rightarrow \infty \).
2. A Pareto distribution: a linear \( e(u) \).
3. An exponential distribution: the constant \( e(u) = 1/\lambda \).
4. Light-tailed distributions: \( e(u) \rightarrow 0, \ u \rightarrow \infty \).
Rough methods for the detection of heavy tails and the number of finite moments.

Plots of function $e(u)$ for several distributions.
4. Hill’s and other estimators of the tail index

The Hill estimate works bad if

1. the underlying DF does not have a regularly varying tail,
2. the tail index $\alpha = 1/\gamma$ is not positive,
3. the sample size is not large enough,
4. the tail is not heavy enough, i.e. $\gamma$ is not big,
5. $F \subseteq R_\alpha$, since it strongly depends on the slowly varying function $\ell(x)$ that is usually unknown.

The disadvantages of Hill’s estimate show that one has to apply several estimates of the tail index to deal with the complex analysis of data.
Definition of the independence.

Random variables $\xi_1, \xi_2, \ldots, \xi_n$ are independent if for any $x_1, x_2, \ldots, x_n$ it holds

\[
P\{\xi_1 \leq x_1, \xi_2 \leq x_2, \ldots, \xi_n \leq x_n\} = P\{\xi_1 \leq x_1\}P\{\xi_2 \leq x_2\} \cdots P\{\xi_n \leq x_n\}
\]

or in terms of cumulative distribution functions

\[
F(x_1, x_2, \ldots, x_n) = F_1(x_1)F_2(x_2) \cdots F_n(x_n),
\]

where $F_k(x_k)$ is a distribution function of the r.v. $\xi_k$. 
Rough methods for the dependence detection.

Correlation between two r.v.s $X_1$ and $X_2$ is defined by

$$\rho(X_1, X_2) = \frac{\text{cov}(X_1, X_2)}{\sqrt{\text{var}(X_1)\text{var}(X_2)}},$$

where

$$\text{cov}(X_1, X_2) = \mathbb{E}(X_1 X_2) - \mathbb{E}(X_1)\mathbb{E}(X_2)$$

is a covariance, and $\text{var}(x)$ is a variance.

1. If $X_1$ and $X_2$ are independent then $\rho(X_1, X_2) = 0$.
2. If $\rho(X_1, X_2) = 0$ then not necessary $X_1$ and $X_2$ are independent!
3. If Gaussian r.v.s $X_1$ and $X_2$ are not correlated then they are independent.
4. For non-Gaussian r.v.s it may be not true.
Rough methods for the dependence detection.

Generally, the covariances and correlations cannot indicate the dependence.

1. They describe the **degree to which two r.v.s. are linearly dependent**:

   \[ \rho(X_1, X_2) \in [-1, 1]. \]

2. \[ \rho(X_1, X_2) = \pm 1 \iff X_1 \text{ and } X_2 \text{ are perfectly linearly dependent, i.e. } X_2 = \alpha + \beta X_1, \text{ for some } \alpha \in R \text{ and } \beta \neq 0. \]

What tool could be an appropriate to indicate the dependence?
Exact methods for the dependence detection.

The mixing conditions of a stationary sequence should be considered for dependence detection.

Definition

(Rosenblatt (1956b)) The strictly stationary ergodic sequence of random vectors $X_t$ is strongly mixing with rate function $\phi_k$ for $\sigma$-field, if

$$\sup_{A \in \sigma(X_t,t \leq 0), B \in \sigma(X_t,t > k)} |P(A \cap B) - P(A)P(B)| = \phi_k \to 0, k \to \infty.$$  

The rate $\phi_k$ shows how fast the dependence between the past and the future decreases.
Another measure of dependence is given by the dependence index:

\[ \beta_n = \sup_{x,y} \left| \sum_{j=1}^{n} f_j(x, y) - f(x)f(y) \right|, \]

where \( f(x) \) is a marginal probability density function (PDF) of a stationary sequence \( \{X_j, j = 1, 2, \ldots\} \), \( f_j(x, y) \) is a joint PDF of \( X_1 \) and \( X_{1+j}, j = 1, 2, \ldots \).

1. for i.i.d. sequences \( \beta_n = 0 \) for all \( n \),
2. for sequences with strong long range dependence \( \beta_n \) may tend to infinity, and
3. in between \( \beta_n \) may converge to a finite limit at various rates.

It is difficult to estimate such dependence measures by statistical tools.
Rough methods for the dependence detection.

**The autocorrelation function (ACF) is**

\[
\rho_X(h) = \rho(X_t, X_{t+h}) = \mathbb{E} \left( (X_t - \mathbb{E}(X_t))(X_{t+h} - \mathbb{E}(X_{t+h})) \right) / \text{Var}(X_t)
\]

Let \( \{X_t, t = 0, \pm 1, \pm 2, \ldots\} \) be a stationary sample series. The standard sample ACF at lag \( h \in \mathbb{Z} \) is

\[
\rho_{n,X}(h) = \frac{\sum_{t=1}^{n-h} (X_t - \bar{X}_n)(X_{t+h} - \bar{X}_n)}{\sum_{t=1}^{n} (X_t - \bar{X}_n)^2},
\]

where \( \bar{X}_n = \frac{1}{n} \sum_{t=1}^{n} X_t \) represents the sample mean.

The accuracy of \( \rho_{n,X}(h) \) may be poor if the sample size \( n \) is small or \( h \) is large. The relevance of \( \rho_{n,X}(h) \) is determined by its rate of convergence to the real ACF. When the distribution of the \( X_t \)'s is very heavy-tailed (in the sense that \( \mathbb{E}X_t^4 = \infty \)), this rate can be extremely slow.
The autocorrelation function for heavy-tailed data.

For heavy-tailed data with infinite variance it is better to use the modified sample ACF:

\[ \tilde{\rho}_{n,X}(h) = \frac{\sum_{t=1}^{n-h} X_t X_{t+h}}{\sum_{t=1}^{n} X_t^2}, \]

i.e. without \( \overline{X}_n \).

However, this estimate may behave in a very unpredictable way if one uses the class of non-linear processes in the sense that this sample ACF may converge in distribution to a non-degenerate r.v. depending on \( h \).
For linear models it converges in distribution to a constant depending on \( h \), Davis and Resnick (1985).
Confidence intervals of the sample ACF: linear processes.

**The causal ARMA process** (autoregressive moving average)

The model ARMA(p,q) has the form

\[ X_t = \sum_{j=0}^{p} \theta_j X_{t-j} + \sum_{j=0}^{q} \psi_j Z_{t-j}, \quad t = 1, \ldots, n. \]

The model MA(q) has the form

\[ X_t = \sum_{j=0}^{q} \psi_j Z_{t-j}. \]
Confidence intervals of the sample ACF: linear processes.

Bartlett’s formula.

If the process is linear (1), \( \sum_{j=-\infty}^{\infty} |\psi_j| < \infty \) and \( \mathbb{E}Z_t^4 < \infty \), then standard sample ACF \( \rho_{n,X}(i) \) has the asymptotic joint normal distribution with mean \( \rho_X(i) \) (i.e. ACF) and variance

\[
\text{var} (\rho_{n,X}(i)) = \frac{c_{ii}}{n},
\]

\[
c_{ii} = \sum_{k=-\infty}^{\infty} \left[ \rho_X^2(k + i) + \rho_X(k - i)\rho_X(k + i) + 2\rho_X^2(i)\rho_X^2(k) - 4\rho_X(i)\rho_X(k)\rho_X(k + i) \right],
\]

as \( n \to \infty \).

The Bartlett’s formula (2) allows to check the hypothesis \( \rho_{n,X}(i) = 0 \).
Confidence intervals of the sample ACF.

Example

For i.i.d. white noise $Z_t$ we have $\rho_Z(0) = 1$ and $\rho_Z(i) = 0$ for $i \neq 0$ (since $Z_t$ and $Z_{t+i}$ are independent) and $\text{var} (\rho_{n,Z}(i)) = 1/n$ by (2).

For ARMA process driven by such a white noise $Z_t$

1. the sample ACF is approximately normally distributed with mean 0 and variance $1/n$ for sufficiently large $n$.
2. It provides 95% confidence interval with the bounds $\pm 1.96/\sqrt{n}$ for the sample ACF.
3. The hypothesis $\rho_{n,X}(i) = 0$ is accepted if $\rho_{n,X}(i)$ falls within this interval.

What would happened if

1. the noise $Z_t$ is not normal, and (or)
2. the process is not linear?

Natalia Markovich  Analysis methods of heavy-tailed data
Confidence intervals of the heavy-tailed sample ACF $\tilde{\rho}_{n,X}(h)$: linear processes.

ARMA process with i.i.d. regularly varying noise and tail index $0 < \alpha < 2$ (infinite variance).

$\tilde{\rho}_{n,X}(h)$ estimates the quantity $\sum_j \psi_j \psi_{j+h} / \sum_j \psi_j^2$ that represents the autocorrelation $\text{cov}(X_0, X_h)$ in the case of a finite variance.

Illusion: heavy-tailed sample ACF $\tilde{\rho}_{n,X}(h)$ can be applied to heavy-tailed processes without problem.


- For $\alpha < 1$ use $\tilde{\rho}_{n,X}(h)$;
- for $1 < \alpha < 2$ the classical sample ACF $\rho_{n,X}(h)$.

The calculation of confidence intervals in both cases is not easy.
Confidence intervals of the heavy-tailed sample ACF $\tilde{\rho}_n, h(t)$: nonlinear processes.

GARCH($p, q$) process (Mikosch (2002)):

$$X_t = \mu + \sigma_t Z_t, \quad t \in \mathbb{Z},$$

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2,$$

- $(Z_t)$ is an i.i.d. sequence,
- $\sigma_t$ and $Z_t$ are independent for a fixed $t$,
- the mean $\mu$ is estimated from the data, particularly $\mu = 0$,
- $\alpha_i$ and $\beta_j$ are non-negative constants.

If the marginal distribution of the time series is very heavy-tailed, namely, the fourth moment is infinite, then the asymptotic normal confidence bounds for the sample ACF are not applicable anymore.
Testing of long-range dependence (LRD).

Definition
A stationary process \((X_t)\) is long range dependent if
\[
\sum_{h=0}^{\infty} |\rho_X(h)| = \infty,
\]
where \(\rho_X(h)\) is the ACF at lag \(h \in \mathbb{Z}\), and short range dependent otherwise.

- This property implies that even though \(\rho_X(h)\)’s are individually small for large lags, their cumulative effect is important.
- To detect LRD by statistical procedures one replaces \(\rho_X(h)\) by the sample ACF \(\rho_{n,X}(h)\).
- The LRD effect is typical for long time series, e.g., several thousand points. One can look at lags 250, 300, 350, etc.
- Usual short range dependent data sets would show a sample ACF dying after only a few lags and then persisting within the 95% Gaussian confidence window \(\pm 1.96/\sqrt{n}\).
Testing of long-range dependence: Hurst parameter.

One can assume that for some constant $c_\rho > 0$

$$\rho_X(h) \sim c_\rho h^{2(H-1)} \quad \text{for large } h \quad \text{and} \quad H \in (0.5, 1),$$

(LRD case).

- The constant $H \in (0.5, 1)$ is called the Hurst parameter.
- The closer $H$ is to 1 the slower is the rate of $\rho_X(h)$ to zero as $h \to \infty$, i.e., the longer is the range of dependence in the time series.
- Kettani and Gubner (2002):

$$\hat{H}_n = 0.5 \left( 1 + \log_2(1 + \rho_n X(1)) \right).$$
Rank coefficient Kendall’s $\tau$ can be estimated by the sample:

$$\rho_\tau = \frac{2S_\tau}{n(n-1)},$$

where

$$S_\tau = \sum_{i=1}^{n} \sum_{j=i+1}^{n} \text{sign} (r_j - r_i),$$

where $r_i$ is the order number of the individuum by the second property with the number $i$ by the first property.

Rank coefficient Spearman’s $\rho$ can be estimated by the sample:

$$\rho_S = 1 - \frac{6S_\rho}{n^3 - n},$$

where

$$S_\rho = \sum_{i=1}^{n} (r_i - i)^2$$

Pickands dependence function $A(t)$.

$\rho_\tau$ and $\rho_S$ can be represented by means of $A(t)$. 

Natalia Markovich

Analysis methods of heavy-tailed data
Let \((X_1, Y_1), \ldots, (X_n, Y_n)\) be a bivariate i.i.d. random sample with a bivariate extreme value distribution \(G(x, y)\).

**Example:** \(X_1\) is a TCP-flow file size, \(Y_1\) is a duration of its transmission.

Similarly to univariate case it implies, that there exist normalizing constants \(a_{j,n} > 0\) and \(b_{j,n} \in R, j = 1, 2\) such that as \(n \to \infty\),

\[
P\left\{ \left( \frac{M_{1,n} - b_{1,n}}{a_{1,n}} \right) \leq x, \left( \frac{M_{2,n} - b_{2,n}}{a_{2,n}} \right) \leq y \right\} \quad (3)
\]

\[
= F^n(a_{1,n}x + b_{1,n}, a_{2,n}y + b_{2,n}) \to G(x, y),
\]

where \(M_{1,n} = \max(X_1, \ldots, X_n), M_{2,n} = \max(Y_1, \ldots, Y_n)\) are the component-wise maxima.

The vector \((M_{1,n}, M_{2,n})\) will in general not be presence in the original data.
Definition

Let $F_1(x)$ and $F_2(y)$ be the DFs of $X$ and $Y$.

Let $G_j(x)$, $j = 1, 2$ be a univariate extreme value DF and $F_j(x)$ is in its domain of attraction.

$G(x, y)$ may be determined by margins $G_1(x)$ and $G_2(y)$ by the representation

$$G(x, y) = \exp \left( \log (G_1(x)G_2(y)) \cdot A \left( \frac{\log (G_2(y))}{\log (G_1(x)G_2(y))} \right) \right),$$

where $A(t)$, $t \in [0, 1]$, is the Pickands dependence function.
In the bivariate case the function $A(t)$ satisfies two properties:

1. $(1 - t) \lor t \leq A(t) \leq 1$, $t \in [0, 1]$, i.e. $A(0) = A(1) = 1$ and lies inside the triangle determined by points $(0, 1)$, $(1, 1)$ and $(0.5, 0.5)$;

2. $A(t)$ is convex.

Case $A(t) \equiv 1$ corresponds to a total independence

and $A(t) = (1 - t) \lor t$ corresponds to a total dependence.
A-estimators

Capéraà et al. (1997):

\[ \hat{A}_{n}^{HT}(t) = \left( \frac{1}{n} \sum_{i=1}^{n} \min \left( \frac{\hat{\xi}_{i} / \bar{\xi}_{n}}{1 - t}, \frac{\hat{\eta}_{i} / \bar{\eta}_{n}}{t} \right) \right)^{-1} \]

Hall and Tajvidi (2000):

\[ \log \hat{A}_{n}^{C}(t) = \frac{1}{n} \sum_{i=1}^{n} \log \max \left( t\hat{\xi}_{i}, (1 - t)\hat{\eta}_{i} \right) \]
\[ -t \frac{1}{n} \sum_{i=1}^{n} \log \hat{\xi}_{i} - (1 - t) \frac{1}{n} \sum_{i=1}^{n} \log \hat{\eta}_{i} \]

Here \( \hat{\xi}_{i} = -\log \hat{G}_{1}(X_{i}) \) and \( \hat{\eta}_{i} = -\log \hat{G}_{2}(Y_{i}) \), \( i = 1, \ldots, n \),
\( \bar{\xi}_{n} = n^{-1} \sum_{i=1}^{n} \hat{\xi}_{i} \), \( \bar{\eta}_{n} = n^{-1} \sum_{i=1}^{n} \hat{\eta}_{i} \).
Problems of $A$-estimators

1. The estimators are not convex. They may be improved by taking a convex hull.

2. The margins $G_1(x)$ and $G_2(x)$ are unknown. One has to replace them by their estimates $\hat{G}_1(x)$ and $\hat{G}_2(x)$, e.g., by empirical DFs constructed by component-wise maxima over blocks of data. The amount of these maxima may be very moderate that may reflect on the accuracy of an empirical DF.

3. The component-wise maxima may be not observable together, i.e. there are such pairs of maxima which do not presence in the sample.
Both estimates of $A(t)$ allow to get the estimate $\hat{G}(x, y)$ and construct bivariate quantile curves

$$Q(\hat{G}, p) = \{(x, y) : \hat{G}(x, y) = p\}, \quad 0 < p < 1, \quad (4)$$

$$\hat{G}(x, y) \approx \exp \left( \log \left( \frac{\hat{G}_1(x) \hat{G}_2(y)}{\hat{A}} \right) \hat{A} \left( \frac{\log \left( \hat{G}_2(y) \right)}{\log \left( \frac{\hat{G}_1(x) \hat{G}_2(y)}{\hat{A}} \right)} \right) \right),$$

assuming that $\hat{G}_1(x) = p^{(1-w)/\hat{A}(w)}$ and $\hat{G}_2(x) = p^{w/\hat{A}(w)}$ for some $w \in [0, 1]$ in order to get $\hat{G}(x, y) = p$, Beirlant et al. (2004).

The quantile curve consists of the points

$$Q(\hat{G}, p) = \{ \left( \hat{G}_1^{-1}(p^{(1-w)/\hat{A}(w)}), \hat{G}_2^{-1}(p^{w/\hat{A}(w)}) \right) : w \in [0, 1] \}.$$
Characteristics of sub-sessions:
- the size of a sub-session (s.s.s);
- the duration of a sub-session (d.s.s.).

Characteristics of the transferred Web-pages:
- the size of the response (s.r.);
- the inter-response time (i.r.t.).
Description of the Web data

Main statistical characteristics of the Web-data.

<table>
<thead>
<tr>
<th></th>
<th>s.s.s.(B)</th>
<th>d.s.s.(sec)</th>
<th>s.r.(B)</th>
<th>i.r.t.(sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Size</td>
<td>373</td>
<td>373</td>
<td>7107</td>
<td>7107</td>
</tr>
<tr>
<td>Minimum</td>
<td>128</td>
<td>2</td>
<td>0</td>
<td>6.543 · 10^{-3}</td>
</tr>
<tr>
<td>Maximum</td>
<td>5.884 · 10^7</td>
<td>9.058 · 10^4</td>
<td>2.052 · 10^7</td>
<td>5.676 · 10^4</td>
</tr>
<tr>
<td>Mean</td>
<td>1.283 · 10^6</td>
<td>1.728 · 10^3</td>
<td>5.395 · 10^4</td>
<td>80.908</td>
</tr>
<tr>
<td>StDev</td>
<td>4.079 · 10^6</td>
<td>5.206 · 10^3</td>
<td>4.931 · 10^5</td>
<td>728.266</td>
</tr>
</tbody>
</table>
Results of the Web traffic analysis.

Estimation of the tail index for Web-traffic characteristics.

<table>
<thead>
<tr>
<th>r.v.</th>
<th>c</th>
<th>$m_b$</th>
<th>$\gamma^b_1$</th>
<th>$\gamma^b_2$</th>
<th>$\gamma^H_{n,k}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>s.s.s.</td>
<td>0.3</td>
<td>8</td>
<td>1.179</td>
<td>0.877</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>10</td>
<td>0.856</td>
<td>0.763</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>22</td>
<td>0.902</td>
<td>0.763</td>
<td></td>
</tr>
<tr>
<td>s.r.</td>
<td>0.3</td>
<td>72</td>
<td>0.75</td>
<td>0.763</td>
<td>0.763</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>71</td>
<td>0.87</td>
<td>0.763</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>92</td>
<td>0.85</td>
<td>0.763</td>
<td></td>
</tr>
<tr>
<td>i.r.t.</td>
<td>0.3</td>
<td>42</td>
<td>0.69</td>
<td>0.495</td>
<td>0.38</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>65</td>
<td>0.625</td>
<td>0.495</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>156</td>
<td>0.611</td>
<td>0.495</td>
<td></td>
</tr>
<tr>
<td>d.s.s.</td>
<td>0.3</td>
<td>10</td>
<td>0.658</td>
<td>0.739</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>13</td>
<td>0.539</td>
<td>0.739</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>18</td>
<td>0.683</td>
<td>0.739</td>
<td></td>
</tr>
</tbody>
</table>
Results of the Web traffic analysis.

Notations to the previous table:

- $\gamma^b_i$ is the group estimator with the bootstrap selected parameter $m (m_b)$;
- $\gamma^p_i$ is the group estimator with the plot selected parameter $m$;
- $\hat{\gamma}^H_{n,k}$ is the Hill’s estimator with the plot selected parameter $k$. 

Natalia Markovich
Analysis methods of heavy-tailed data
The $EVI$ estimation by the Hill’s estimator and the group estimator $\gamma_l$ for the data sets size of sub-sessions (left) and duration of sub-sessions (right).
The $EVI$ estimation by the Hill's estimator and the group estimator $\gamma_l$ for the data sets inter-response times (left) and size of responses (right).
Conclusions from analysis of the tail index:

1. the distributions of considered Web-traffic characteristics are heavy-tailed;

2. at least $\beta$th moments, $\beta \geq 2$ of the distribution of the s.s.s., s.r., d.s.s. are not finite;

3. the distribution of i.r.t. has two finite moments;

4. it might be possible for s.s.s. (when $1 < \hat{\gamma} < 2$) that $\alpha < 1$ and the expectation could be also not finite.
Results of the Web traffic analysis.

Analysis of values $R_n(p)$.

- The values $R_n(p)$ are dramatically large for large $n$ and $p \geq 2$, e.g., in the case of the duration of sub-sessions and $n = 350 \ln R_n(p) \approx 10$ for $p = 2$ and $\ln R_n(p) \approx 10^3$ for $p = 3$.

- Hence, one may conclude that all moments of the considered r.v.s apart from the first one are not finite.
\( n \rightarrow \ln R_n(p) \) of the duration of sub-sessions (left) and the size of sub-sessions (right) for a variety of \( p \)-values.
$n \rightarrow \ln R_n(p)$ of the inter-response times (left) and the size of responses (right) for a variety of $p$-values.
Results of the Web traffic analysis.

Analysis of the mean excess function.

- The plots $u \to e_n(u)$ tend to infinity for large $u$ implying heavy tails.

- These plots are close to a linear shape for all sets of data. The latter implies that the considered distributions can be modelled by a $DF$ of a Pareto type.
Exceedance $e_n(u)$ against the threshold $u$ for the duration of sub-sessions (left) and the size of sub-sessions (right).
Exceedance $e_n(u)$ against the threshold $u$ for the inter-response times (left) and the size of responses (right).
Results of the Web traffic analysis.

Analysis of QQ-plots.

- Two versions of QQ-plots are shown.
- Right top plots show that the distribution of the d.s.s. is close to lognormal and Pareto distributions; distributions of s.s.s., i.r.t. and s.r. are close to Pareto.
- Left top plots show that the exponential distribution cannot be used as an appropriate model for these r.v.s.; distributions of d.s.s., s.s.s., i.r.t. and s.r. are close to a Generalized Pareto distribution with the DF

$$\Psi_{\sigma,\gamma}(x) = \begin{cases} 
1 - (1 + \gamma x / \sigma)^{-1/\gamma}, & \gamma \neq 0, \\
1 - \exp(-x / \sigma), & \gamma = 0,
\end{cases}$$

with different values of the parameters $\gamma$ and $\sigma$.
- The QQ-plot does not give a unique model to fit the underlying distribution.
QQ-plots for the duration of sub-sessions.

Left: exponential quantiles (top) and quantiles of the GPD(0.3;1) (bottom) against d.s.s./s (the linear curves correspond to the QQ-plot of the same distributions).

Right: Empirical DFs of the r.v. $U_i = F(X_i)$. Exponential, Pareto, Weibull, lognormal and normal distributions are used as models of $F$. The linear curve corresponds to the exponential distribution.
QQ-plot for the size of sub-sessions.

Left: exponential quantiles (top) and quantiles of the GPD(1;0.015) (bottom, first plot) and the GPD(0.05;0.3) (bottom, second plot) against s.s.s./s (the linear curves correspond to the QQ-plot of the same distributions).

Right: Empirical DFs of the r.v. $U_i = F(X_i)$. Exponential, Pareto, Weibull, lognormal and normal distributions are used as the models of $F$. The linear curve corresponds to the exponential distribution.
QQ-plots for the inter-response times.

Left: exponential quantiles (top) and quantiles of the GPD(0.8;0.015) (bottom) against i.r.t./s (the linear curves correspond to the QQ-plot of the same distributions).

Right: Empirical DFs of the r.v. $U_i = F(X_i)$. Exponential, Pareto, Weibull, lognormal and normal distributions are used as models of $F$. The linear curve corresponds to the exponential distribution.
QQ-plot for the size of responses.

**Left:** exponential quantiles (top) and quantiles of the GPD(1;0.015) (bottom) against s.r./s (the linear curves correspond to the QQ-plot of the same distributions).

**Right:** Empirical DFs of the r.v. $U_i = F(X_i)$. Exponential, Pareto, Weibull, lognormal and normal distributions are used as the models of $F$. The linear curve corresponds to the exponential distribution.
Results of the Web traffic analysis.

**ACF analysis of Web-data.**

The previous analysis shows that

- the considered Web data are **heavy-tailed with infinite variance**. Therefore, the application of formula

\[
\tilde{\rho}_{n,\chi}(h) = \frac{\sum_{t=1}^{n-h} X_t X_{t+h}}{\sum_{t=1}^{n} X_t^2}
\]

is relevant.
### Summary results of the preliminary analysis

**Comparison of the recommended methods for Web traffic data**

<table>
<thead>
<tr>
<th>r.v.</th>
<th>Amount of finite moments $R_n(p)$</th>
<th>Hill &amp; Group estimator</th>
<th>Type of distribution QQ-plot $e_n(u)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>s.s.s. (B)</td>
<td>1</td>
<td>1</td>
<td>$GPD(0.015; 1)$, $GPD(0.05; 0.3)$</td>
</tr>
<tr>
<td>d.s.s. (sec)</td>
<td>1</td>
<td>1</td>
<td>$GPD(1; 0.3)$, lognormal</td>
</tr>
<tr>
<td>s.r. (B)</td>
<td>1</td>
<td>1</td>
<td>$GPD(0.015; 1)$</td>
</tr>
<tr>
<td>i.r.t. (sec)</td>
<td>1</td>
<td>2</td>
<td>$GPD(0.015; 0.8)$</td>
</tr>
</tbody>
</table>
Testing of dependence

ACF estimation by the modified sample ACF and the standard sample ACF for the data sets s.s.s. (first two plots left), d.s.s. (last two plots right).

The dotted horizontal lines indicate 95% asymptotic confidence bounds ($\pm 1.96/\sqrt{n}$) corresponding to the ACF of i.i.d. Gaussian r.v.s.
Testing of dependence

ACF estimation by the modified sample ACF and the standard sample ACF for the data sets i.r.t. (first two plots left), s.r. (last two plots right).

The dotted horizontal lines indicate 95% asymptotic confidence bounds \((\pm 1.96/\sqrt{n})\) corresponding to the ACF of i.i.d. Gaussian r.v.s.
Hurst parameter estimation for Web traffic data

Table: Hurst parameter estimation for Web traffic.

<table>
<thead>
<tr>
<th>Data</th>
<th>s.s.s.</th>
<th>d.s.s.</th>
<th>i.r.t.</th>
<th>s.r.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{H}_n$</td>
<td>0.493</td>
<td>0.488</td>
<td>0.508</td>
<td>0.507</td>
</tr>
</tbody>
</table>

Main conclusion:

all data sets are heavy-tailed and not long-range dependent.
TCP-flow analysis

We observe

- TCP-flow sizes $S$ and
- durations $D$

gathered from one source destination pair.

Motivation is to estimate

- the distribution of the maximal rate (or throughput) $R = S/D$ and
- the expected throughput $R$ (or $S/D$)

that the transport system provides.
Description of the TCP-flow data

The analyzed data consist of:

- TCP-flow sizes and durations of transmissions have been measured from the mobile network of the Finnish operator Elisa;
- mobile TCP connections from periods of low, average and high network load conditions;
- TCP flows on port 80 (a WWW (HTTP) application).
- The number of analyzed flows is 610 000 and, for practical reasons, we consider 61 disjoint bivariate samples, each of size $n = 10 000$. 
## Description of the TCP-flow data

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Unit</th>
<th>Definition</th>
<th>Sample Mean</th>
<th>Sample Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>kB</td>
<td>Content</td>
<td>9.0</td>
<td>1303</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Transmitted</td>
<td>9.5</td>
<td>1357</td>
</tr>
<tr>
<td>Duration</td>
<td>sec</td>
<td>SYN-FIN</td>
<td>18.2</td>
<td>2219</td>
</tr>
</tbody>
</table>

The results, [min, max] ranges over all 61 samples.

'Content' refers to the size of the downloaded web content and 'Transmitted' means Content plus segments retransmitted by TCP. Both are measures of the size of a flow. 'SYN-FIN' means from the three-way handshaking (synchronization) to finish.
Table: Estimation of the EVI for flow sizes ("Content" and "Transmitted") and durations ("SYN-FIN")

<table>
<thead>
<tr>
<th></th>
<th>$\hat{\gamma}^H(n, k)$</th>
<th>$\gamma_I$</th>
<th>$\hat{\gamma}^M(n, k)$</th>
<th>$\hat{\gamma}^{UH}(n, k)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Min</td>
<td>Max</td>
<td>Min</td>
<td>Max</td>
</tr>
<tr>
<td>Content</td>
<td>0.59</td>
<td>0.87</td>
<td>0.45</td>
<td>0.98</td>
</tr>
<tr>
<td>Transmitted</td>
<td>0.58</td>
<td>1.15</td>
<td>0.45</td>
<td>0.94</td>
</tr>
<tr>
<td>SYN-FIN</td>
<td>0.52</td>
<td>1.00</td>
<td>0.38</td>
<td>0.77</td>
</tr>
</tbody>
</table>
Conclusions from analysis of the tail index for TCP-flow data:

1. the distributions of TCP-flow size and duration are heavy-tailed;

2. all estimators apart of the group estimator $\gamma_l$ indicate that the flow sizes samples (both content and transmitted) may have infinite variance under the assumption that their distributions are regularly varying;

3. some samples of flow durations may have two finite first moments.
**Table: Comparison of the “rough” methods for TCP-flow data**

<table>
<thead>
<tr>
<th></th>
<th>Amount of first finite moments</th>
<th>Type of distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R_n(p)$</td>
<td>QQ-plot</td>
</tr>
<tr>
<td>Content</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Transmitted</td>
<td>1</td>
<td>&lt; 1</td>
</tr>
<tr>
<td>SYN-FIN</td>
<td>&lt; 1</td>
<td>&lt; 1</td>
</tr>
</tbody>
</table>
Testing of dependence

ACF estimation by the modified sample ACF and the standard sample ACF of one sub-sample \((n = 1000)\) of the TCP-flow sizes (first two plots left), and durations (last two plots right).

The horizontal lines indicate 95\% asymptotic confidence bounds \(\left(\pm 1.96 / \sqrt{n}\right)\).

Conclusions:

- The TCP-flow sizes may be independent.
- The ACFs of the TCP-flow durations have three clusters that may indicate the dependence.
Testing of long-range dependence

Table: Hurst parameter estimation for Web traffic and TCP-flow data

<table>
<thead>
<tr>
<th>Data</th>
<th>TCP – flow size</th>
<th>TCP – flow duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{H}_n$</td>
<td>0.498</td>
<td>0.506</td>
</tr>
</tbody>
</table>

Main conclusion:
TCP-flow size and duration data sets are heavy-tailed and not long-range dependent.
The distributions of both $S$ and $D$ are heavy-tailed and their expectations may not be finite. Thus, $ES/D$ may be not computable.

Since $S$ and $D$ are dependent and positive, then the DF of the ratio $R = S/D$ is defined by

$$F_R(x) = P\{S/D \leq x\} = \int_0^\infty \int_0^{zx} f(y, z) dy dz$$

$$= \int_0^\infty \int_0^{zx} dF(y, z),$$

$f(y, z)$ is a joint PDF of $S$ and $D$, and its expectation by

$$ER = \int_0^\infty xdF_R(x),$$

if the latter integral converges.
First, we check the dependence between the pairs \((S_1, D_1), \ldots, (S_n, D_n)\) to apply (3). For this purpose, we can calculate the ACF of the r.v.s
\[
 r_i = \sqrt{S_i^2 + D_i^2}, \ i = 1, \ldots, n.
\]
The sample ACF of \(\{r_i\}\) is small in absolute value at all lags (possible exception are two lags that do not persist within 95\% confidence interval). One may suppose that the sizes-duration pairs are independent.
Bivariate analysis of TCP-flow data

Scatter plot of pairs of block maxima \((M_{S,m}^j, M_{D,m}^j)\), \(j = 1, \ldots, 610\), when the block size is \(m = 1000\). The pairs that are presented in the initial sample are marked by dots as far as the pairs that are not presented (e.g., the maximal size in the group does not necessarily correspond to the maximal duration in this group) by circles. Lines \(D = S/384\) and \(D = S/42\) indicate 384 kb/s (EDGE) and 42 kb/s (GPRS) access rates.
Estimates of standard sample ACF of the both maxima samples of size 61 corresponding to TCP-flow sizes (left), and durations (right). The dotted horizontal lines indicate 95% asymptotic confidence bounds $\pm 1.96/\sqrt{n}$. 
Estimates of standard sample ACF of the both maxima samples of size 610 corresponding to TCP-flow sizes (left), and durations (right). The dotted horizontal lines indicate 95% asymptotic confidence bounds $\pm 1.96/\sqrt{n}$. 
The Generalized Extreme Value (GEV) distribution

\[ H_\gamma(x) = \begin{cases} 
\exp\left(-\left(1 + \gamma \left(\frac{x-\mu}{\sigma}\right)^{-1}/\gamma\right)\right), & \gamma \neq 0 \\
\exp\left(-e^{-\left(\frac{x-\mu}{\sigma}\right)}\right), & \gamma = 0.
\end{cases} \]

is applied as a model of the block maxima distribution.

Maximum likelihood estimates of GEV parameters by block maxima of size 610 of TCP-flow data

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Definition</th>
<th>(\gamma)</th>
<th>(\mu)</th>
<th>(\sigma)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>Content</td>
<td>0.332259</td>
<td>7075.92</td>
<td>4605.53</td>
</tr>
<tr>
<td>Duration</td>
<td>SYN-FIN</td>
<td>0.10263</td>
<td>3775.80</td>
<td>2433.27</td>
</tr>
</tbody>
</table>
QQ-plots of block maxima samples

corresponding to TCP-flow sizes (left) and durations (right).
Testing of dependence of the TCP-flow data

The estimation of the Pickands dependence function by estimators $\hat{A}_n^C(t)$ (dashed line) and $\hat{A}_n^{HT}(t)$ (solid line). The maxima sample of size 61 (left) and of size 610 (right). The marginal distributions $G_1(x)$ and $G_2(x)$ of TCP-flow sizes and durations are estimated by GEV.

Conclusions: TCP-flow size and duration are dependent.
Bivariate quantile curves of the TCP-flow data

Using estimates of $A(t)$ one can construct bivariate quantile curves of TCP-flow data by (4).

Estimated quantile curves of TCP-flow data

for $\rho \in \{0.75, 0.9, 0.95\}$ corresponding to estimator $\hat{A}_n^C(t)$: the maxima sample of size 61 (left), of size 610 (right).
Conclusions from bivariate analysis of TCP-flow data

- The analysis is made from samples of moderate size.
- Size $S$ and duration $D$ are heavy-tailed with probably infinite second moment.
- Their distributions are complicated in the sense that they do not belong to any known parametric models.
- Estimates of the Pickands dependence function show that $S$ and $D$ are dependent.
- Bivariate quantile curves show that the bivariate extreme value distribution of $(S, D)$ is 'not quite heavy-tailed' in the sense that not many observations fall in the "outliers area", i.e. beyond the 97.5% quantile curve. This can be a special property of this mobile TCP data.
- Bivariate quantile curves are sensitive to, at least, 
  1. estimation of parameters of margins of $G(x, y)$ and estimates of $A(t)$ and 
  2. the amount of component-wise maxima, or to the block size.
Considering the real data three items have to be investigated:

1. the preliminary detection of heavy tails;
2. the dependence structure of univariate data;
3. the dependence structure of multivariate data.


