Analysis methods of heavy-tailed data

Natalia Markovich

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

July 12, 2022

Modul 3

Statistical analysis

of real data: detection of heavy tails, the number of finite moments and dependence.

The Modul 3 contains

analysis of two kind of telecommunication data: Web traffic data and data of TCP-flows.

Statistical analysis of real data of Web traffic.

Web-traffic analysis

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.).

Main statistical characteristics of the Web-data.

	s.s.s.(B)	d.s.s.(sec)	s.r.(B)	i.r.t.(sec)
Sample Size	373	373	7107	7107
Minimum	128	2	0	$6.543 \cdot 10^{-3}$
Maximum	$5.884 \cdot 10^{7}$	$9.058 \cdot 10^{4}$	$2.052 \cdot 10^{7}$	$5.676 \cdot 10^{4}$
Mean	$1.283 \cdot 10^{6}$	$1.728 \cdot 10^{3}$	$5.395 \cdot 10^{4}$	80.908
StDev	$4.079 \cdot 10^{6}$	$5.206 \cdot 10^{3}$	$4.931 \cdot 10^{5}$	728.266
Scale, s	10^{7}	10^{3}	10^{6}	10^{3}

Results of the Web traffic analysis.

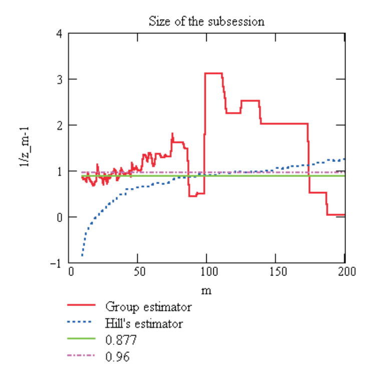
Notations to the next table:

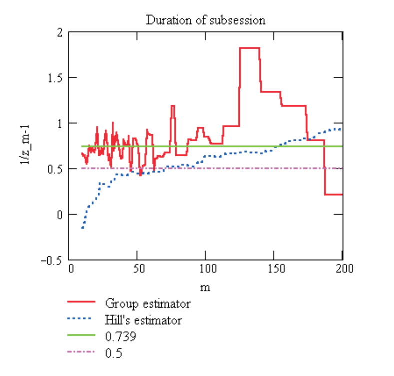
- γ_l^b is the group estimator with the bootstrap selected parameter m (m_b);
- γ_l^p is the group estimator with the plot selected parameter m;
- $\hat{\gamma}_{n,k}^H$ is the Hill's estimator with the plot selected parameter k.
- m_b is the bootstrap-selected parameter m, the group size.
- c is the bootstrap parameter to select m.

Results of the Web traffic analysis.

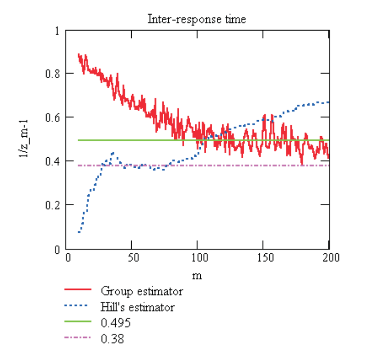
Estimation of the tail index for Web-traffic characteristics.

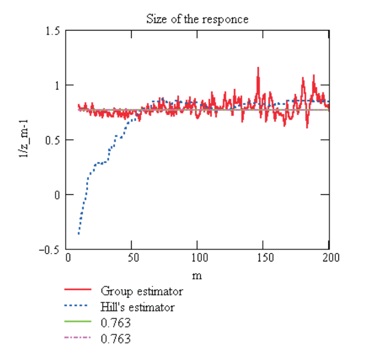
r.v.	c	m_b	γ_l^b	γ_l^p	$\hat{\gamma}_{n,k}^H$
s.s.s.	0.3	8	1.179	0.877	0.96
	0.4	10	0.856		
	0.5	22	0.902		
s.r.	0.3	72	0.75	0.763	0.763
	0.4	71	0.87		
	0.5	92	0.85		
i.r.t.	0.3	42	0.69	0.495	0.38
	0.4	65	0.625		
	0.5	156	0.611		
d.s.s.	0.3	10	0.658	0.739	0.5
	0.4	13	0.539		
	0.5	18	0.683		





The *EVI* estimation by the Hill's estimator and the group estimator γ_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 γ_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 βth 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 $\hat{\alpha} = 1/\hat{\gamma} < 1$ and the expectation could be also not finite.

Results of the Web traffic analysis.

Let $X_1, ..., X_n$ be a sample under study.

$$R_n(p) = M_n(p)/S_n(p), \quad n \geq 1, \quad p > 0$$

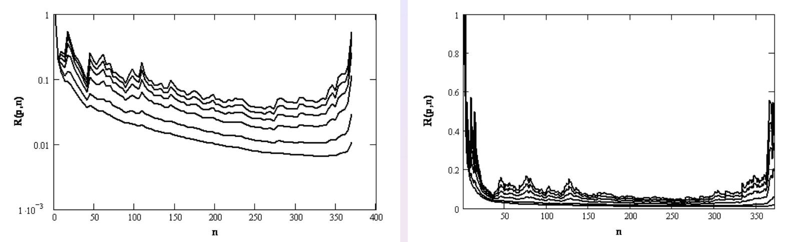
$$S_n(p) = |X_1|^p + \ldots + |X_n|^p$$

,

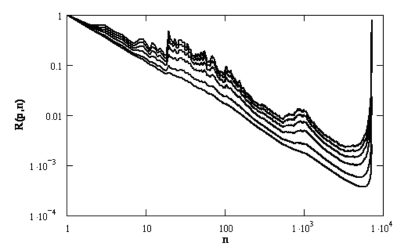
$$M_n(p) = \max\left(|X_1|^p, \ldots, |X_n|^p\right)$$

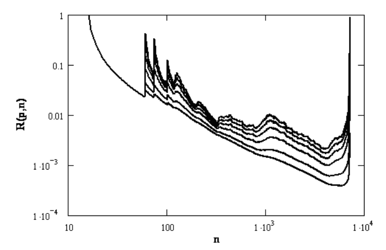
Analysis of values $R_n(p)$

- 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 $p = 3$.
- One may conclude that all moments of order $p = 0.5, 1, 2, 3, 4, 5$ of the considered r.v.s apart from the duration of sub-sessions one are not finite.



$n \to \ln R_n(p)$ **of the duration of sub-sessions (left) and the size of sub-sessions (right) for a variety of** p**-values: curves corresponding to** $p = 0.5, 1, 2, 3, 4, 5$ **are located from bottom to top, respectively.**





$n \to \ln R_n(p)$ **of the inter-response times (left) and the size of responses (right) for a variety of** p**-values: curves corresponding to** $p = 0.5, 1, 2, 3, 4, 5$ **are located from bottom to top, respectively.**

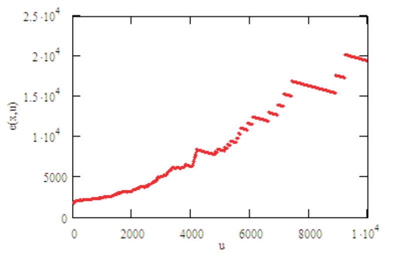
Results of the Web traffic analysis.

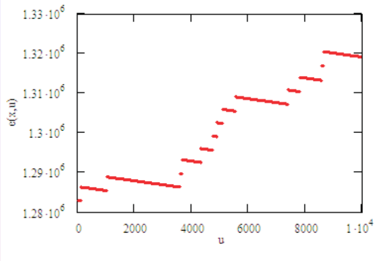
$$e_n(u) = \sum_{i=1}^{n}(X_i - u)\mathbf{1}\{X_i > u\} / \sum_{i=1}^{n}\mathbf{1}\{X_i > u\},$$

u denotes a threshold value.

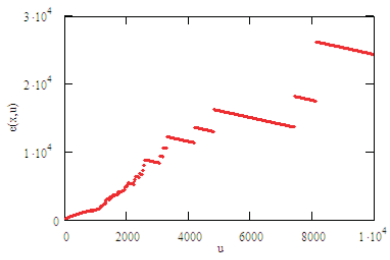
Analysis of the mean excess function $e_n(u)$

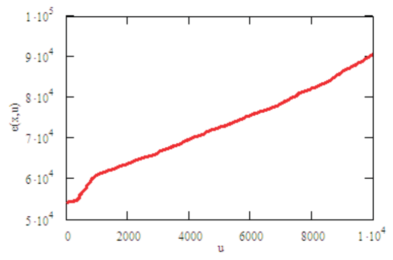
- The plots $u \to e_n(u)$ tend to infinity for large u implying heavy tails.
- These plots are close to a linear shape for all sets of data. The latter implies that the considered distributions can be modelled by a *DF* of a Pareto type.





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).

Analysis of QQ-plots.

- The following model distributions are examine:

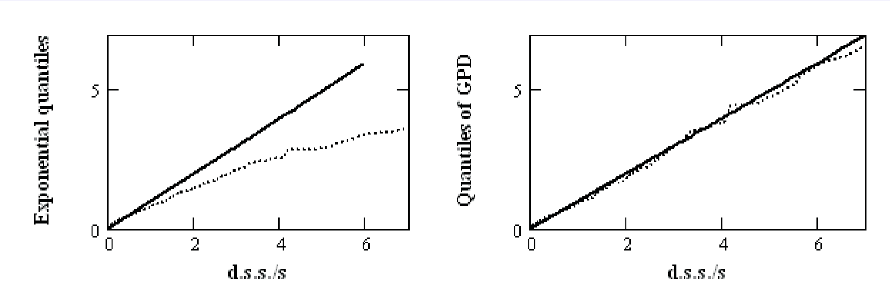
  an exponential distribution,
  a Generalized Pareto distribution with the *DF*

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

  with different values of the parameters γ and σ.

- The QQ-plot does not give a unique model to fit the underlying distribution.

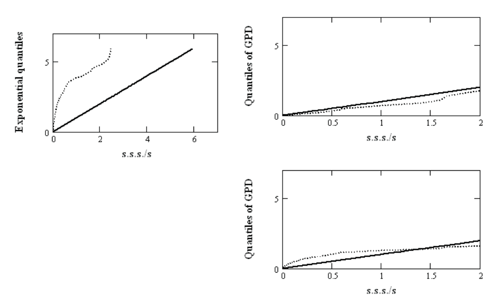
QQ-plots of the duration of sub-sessions



QQ-plots for the duration of sub-sessions (d.s.s./s)

against exponential quantiles (left) and quantiles of GPD(0.3;1) distribution (right) (the linear curves correspond to the appropriate distribution models).
The distribution of the d.s.s. is close to Generalized Pareto distribution.

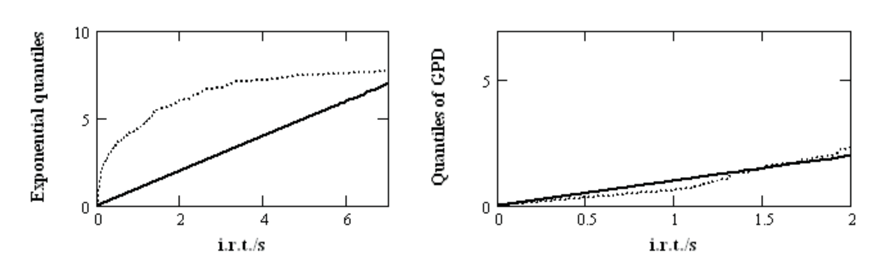
QQ-plots of the size of sub-sessions



QQ-plots for the size of sub-sessions (s.s.s./s)

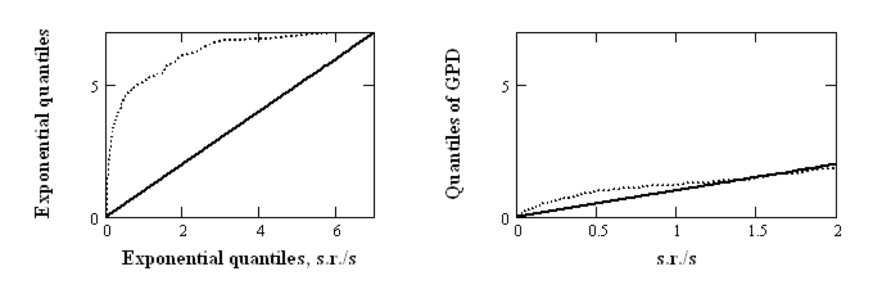
against exponential quantiles (left) and quantiles of GPD(0.015;1) (top right) and GPD(0.05;0.3) distributions (bottom right).
The distribution of the s.s.s. is close to Generalized Pareto distribution.

QQ-plots of the inter-response times



QQ-plots for the inter-response times (i.r.t./s)

against exponential quantiles (left) and quantiles of the GPD(0.015;0.8) distribution (right).
The distribution of the i.r.t. is close to Generalized Pareto distribution.



QQ-plots for the size of responses (s.r./s)

against exponential quantiles (left) and quantiles of the GPD(0.015;1) distribution (right).
The distribution of the s.r. is close to Generalized Pareto distribution.

Summary results of the preliminary analysis

Comparison of the recommended methods for Web traffic data

r.v.	Amount of finite moments $R_n(p)$	Amount of finite moments: Hill & Group estimator	Type of distribution QQ-plot	Type of distribution $e_n(u)$
s.s.s. (B)	1	1	$GPD(0.015; 1)$ $GPD(0.05; 0.3)$	Pareto -like
d.s.s. (sec)	1	1	$GPD(1; 0.3)$, lognormal	Pareto -like
s.r. (B)	1	1	$GPD(0.015; 1)$	Pareto -like
i.r.t. (sec)	1	2	$GPD(0.015; 0.8)$	Pareto -like

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

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

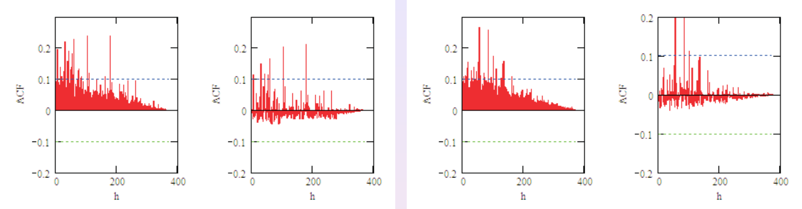
  is relevant.

The standard sample *ACF* at lag $h \in Z$ is

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

$\overline{X}_n = \frac{1}{n}\sum_{t=1}^{n} X_t$ represents the sample mean.

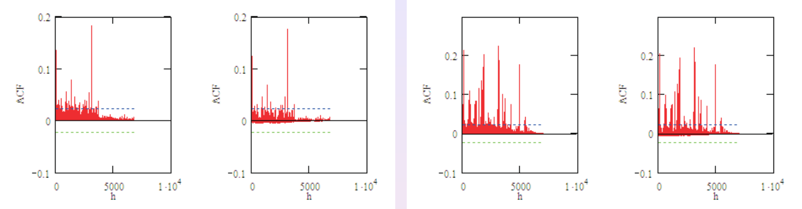
Testing of dependence



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.

Data	$s.s.s.$	$d.s.s.$	$i.r.t.$	$s.r.$
$\hat{H}_n$	0.493	0.488	0.508	0.507

Method by Kettani & Gubner (2002) is used:

$$\hat{H}_n = 0.5\left(1 + \log_2(1 + \rho_{n,X}(1))\right),$$

$\rho_{n,X}(h)$ is sample ACF at lag h.
The closer $H \in (0.5, 1)$ is to 1 the longer is the range of dependence in the time series.

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 $\mathbb{E}R$ (or $\mathbb{E}S/\mathbb{E}D$)

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

Statistic	Unit	Definition	Sample Mean		Sample Variance	
			Min	Max	Min	Max
Size	kB	Content	9.0	20.3	1303	204553
		Transmitted	9.5	20.1	1357	206658
Duration	sec	SYN-FIN	18.2	30.4	2219	52125

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.

Results of the TCP-flow data analysis

Table: Estimation of the EVI γ for flow sizes ("Content" and "Transmitted") and durations ("SYN-FIN")

	$\widehat{\gamma}^H(n,k)$		γ_l		$\hat{\gamma}^M(n,k)$		$\hat{\gamma}^{UH}(n,k)$	
	Min	Max	Min	Max	Min	Max	Min	Max
Content	0.59	0.87	0.45	0.98	0.51	0.98	0.52	0.99
Transmitted	0.58	1.15	0.45	0.94	0.52	0.96	0.53	0.97
SYN-FIN	0.52	1.00	0.38	0.77	0.36	0.86	0.37	0.82

Conclusions from analysis of the tail index for TCP-flow data:

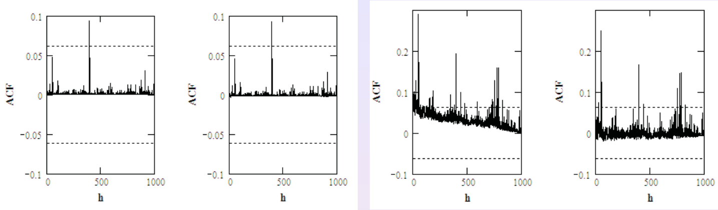
1. the distributions of TCP-flow size and duration are heavy-tailed;
2. all estimators apart of the group estimator γ_l indicate that the flow sizes samples (both content and transmitted) may have infinite variance under the assumption that their distributions are regularly varying;
3. some samples of flow durations may have two finite first moments.

Results of the TCP-flow data analysis

Table: Comparison of the “rough” methods for TCP-flow data

	Amount of first finite moments				Type of distribution	
	$R_n(p)$	Estimators of γ			QQ-plot	$e_n(u)$
Content	1	2	or	1	$GPD(1, 1.3)$	Pareto-like
Transmitted	1	2	or	< 1	$GPD(1, 1.3)$	Pareto-like
SYN-FIN	< 1	2	or	< 1	$GPD(1, 0.85)$	Pareto-like

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 ($\pm 1.96/\sqrt{n}$).

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 TCP-flow data

Data	$TCP - flow\ size$	$TCP - flow\ duration$
$\hat{H}_n$	0.498	0.506

Main conclusion:

TCP-flow size and duration data sets are heavy-tailed and not long-range dependent.

Problems of throughput investigation

- The distributions of both S and D are heavy-tailed and their expectations may not be finite. Thus, $\mathbb{E}S/\mathbb{E}D$ may be not computable.
- Since S and D are dependent and positive, then the DF of the ratio $R = S/D$ is defined by

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

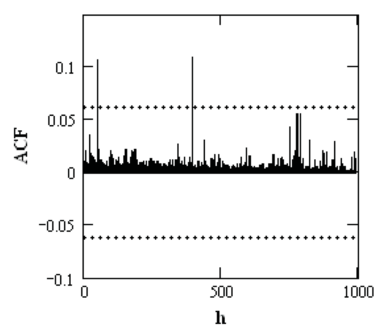
$$= \int_0^{\infty} \int_0^{zx} dF(y,z),$$

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

$$\mathbb{E}R = \int_0^{\infty} x dF_R(x),$$

if the latter integral converges.

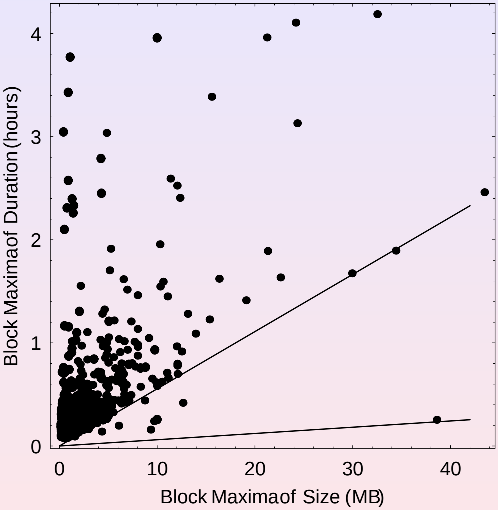
Bivariate analysis of TCP-flow data



First, we check the dependence between the pairs

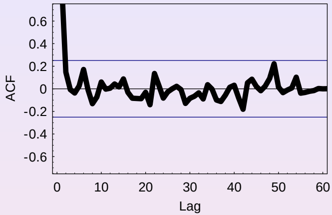
$(S_1, D_1), ..., (S_n, D_n)$ to apply (5) in Modul 2 Lesson 7.
For this purpose, we can calculate the ACF of the r.v.s $r_i = \sqrt{S_i^2 + D_i^2}, i = 1, ..., n$.
The sample ACF of $\{r_i\}$ is small in absolute value at all lags (possible exception are two lags that do not persist within 95% confidence interval). One may suppose that the sizes-duration pairs are independent.

Bivariate analysis of TCP-flow data



Scatter plot of pairs of block maxima $(M^j_{S,m}, M^j_{D,m})$,

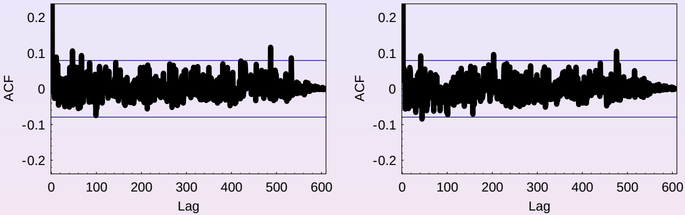
$j = 1, \ldots, 610$, when the block size is $m = 1\,000$.
Lines $D = S/384$ and $D = S/42$ indicate 384 kb/s (EDGE) and 42 kb/s (GPRS) access rates.



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 $\pm 1.96/\sqrt{n}$.

Testing of dependence of the block maxima



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 $\pm 1.96/\sqrt{n}$.

Testing of distribution of the block maxima

The Generalized Extreme Value (GEV) distribution

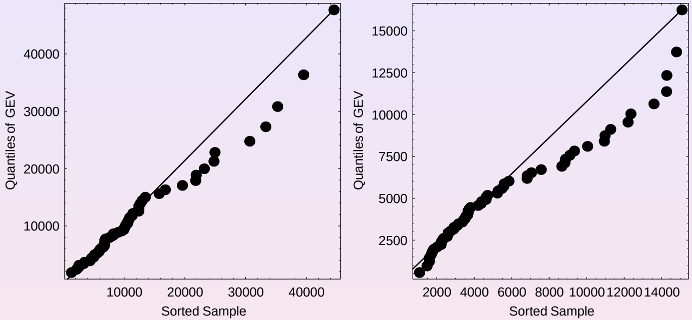
$$H_\gamma(x) = \begin{cases} \exp\left(-\left(1+\gamma\left(\frac{x-\mu}{\sigma}\right)\right)^{-1/\gamma}\right), & \gamma \neq 0 \\ \exp\left(-e^{-\left(\frac{x-\mu}{\sigma}\right)}\right), & \gamma = 0. \end{cases}$$

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

Statistic	Definition	γ	μ	σ
Size	Content	0.332259	7075.92	4605.53
Duration	SYN-FIN	0.10263	3775.8	2433.27

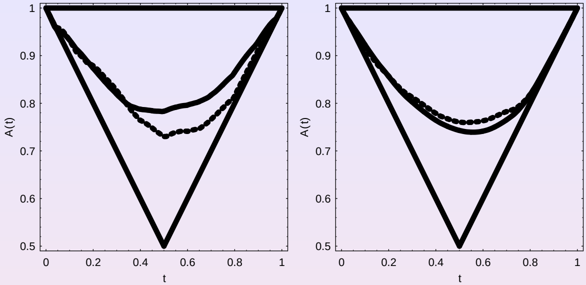
Testing of distribution of the block maxima



QQ-plots of block maxima samples

corresponding to TCP-flow sizes (left) and durations (right).

Testing of dependence of the TCP-flow data

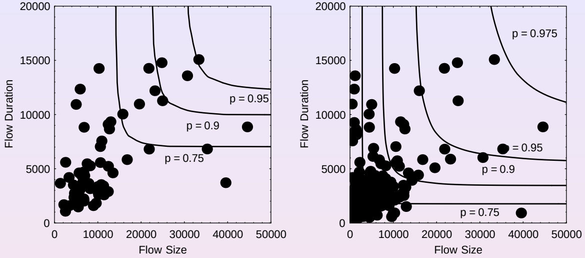


The estimation of the Pickands dependence function

by estimators $\widehat{A}_n^C(t)$ (dashed line) and $\widehat{A}_n^{HT}(t)$ (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 (6) Modul 2, Lesson 7.

Estimated quantile curves of TCP-flow data for $p \in \{0.75, 0.9, 0.95\}$ corresponding to estimator $\widehat{A}_n^C(t)$: 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.

Notes

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.

Reference:

1. **Beirlant, J., Goegebeur, Y., Teugels, J. and Segers, J. (2004)** *Statistics of Extremes: Theory and Applications*. Wiley, Chichester, West Sussex.
2. **Brockwell P.J. and Davis R.A. (1991)** *Time series: Theory and Methods*, 2nd edition. Springer, New York.
3. **Davis, R. and Resnick, S. (1985)** Limit theory for moving averages of random variables with regularly varying tail probabilities. *Ann. Probability* 13, 179-195.
4. **Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997)** *Modeling Extremal Events*. Springer, Berlin.
5. **Hall, P. (1990).** Using the Bootstrap to Estimate Mean Squared Error and Select Smoothing Parameter in Nonparametric Problems. *Journal of Multivariatte Analysis*, 32, 177-203
6. **Kettani, H. and Gubner, J.A. (2002)** A novel approach to the estimation of the hurst parameter in self-similar traffic. In: *Proceedings of IEEE conference on local computer*

Reference:

7. **Markovich, N.M. and Krieger, U.R. (2005)** Statistical Inspection and Analysis Techniques for Traffic Data Arising from the Internet, *HETNETs'04 special journal issue on "Convergent Multi-Service Networks and Next Generation Internet: Performance Modelling and Evaluation"* pp.72/1-72/9.
8. **Markovich, N.M. (2007)** *Nonparametric Analysis of Univariate Heavy-Tailed data: Research and Practice*. Wiley, Chichester, West Sussex.
9. **Resnick, S.I. (2006)** *Heavy-Tail Phenomena. Probabilistic and Statistical Modeling*. Springer, New York.