Department for Applied Statistics
Johannes Kepler University Linz

IFAS Research Paper Series
2009-43

On robust and distribution sensitive Hill like method

Zdeněk Fabián[a] and Milan Stehlík

April 2009

[a]Institute of Computer Science, Academy of Sciences of the Czech Republic

Abstract

In many areas, such as telecommunications and finance, the Pareto approximation to heavy tailed data is too vague. In this paper we introduce a distribution sensitive Hill-like estimator, so called t-Hill estimator. We show that t-Hill estimator for the Pareto distribution is consistent and we demonstrate the robustness of the introduced estimator on both simulated and real data sets.

Keywords:consistency; Hill estimator; insurance; Pareto tail; t-Hill estimator; robustness

1 Introduction

The Pareto-type distribution means that as $x \to \infty$, then survival function $\bar{F}(x) = 1 - F(x)$, where F is the c.d.f., can be written as $\bar{F}(x) = x^{-\alpha} l(x)$, where $\alpha > 0$ and l is a slowly varying function. The parameter $\gamma = 1/\alpha$ is known as the extreme value index or tail index, which helps to indicate the size and frequency of extreme events under F.

Let $X_1, ..., X_n$ be *iid* sample from F. If F is strictly Pareto, $\bar{F}(x) = cx^{-\alpha}, x > x_c$, the distribution of relative excesses $Y_i = X_i/t$ over high threshold t conditionally on $X_i > t$ is Pareto with parameter α and support $[1, \infty)$. Denoting the corresponding order statistics by $X_{1,n} \leq ... \leq X_{n,n}$, Hill (1975) suggested to estimate $\hat{\gamma}$ by

$$\hat{\gamma}_k = H_{k,n} = \frac{1}{k} \sum_{i=1}^{k} \log \frac{X_{n-j+1,n}}{X_{n-k,n}} \tag{1}$$

where $X_{n-k,n}$ is the k-th threshold. The Hill estimator is based on a fact that for a sample $Y_1, ..., Y_n$ from strict Pareto distribution with support $[1, \infty)$ and survival function $\bar{F}(x) = x^{-\alpha}$,

$$\frac{1}{\hat{\alpha}_n} = \frac{1}{n} \sum_{i=1}^{n} \log Y_i$$

is the maximum likelihood estimator of $1/\alpha$. The Hill estimator $H_{k,n}$ was shown by Mason (82) to be consistent estimator for γ (as $k, n \to \infty, k/n \to 0$) whatever the slowly varying function l may be. Since for every choice of k, one obtains another estimator $\hat{\gamma}_k = H_{n,k}$, results are studied by means of Hill plots $\{k, H_{n,k}\}$ for some range of $k \leq n - 1$. However, maximum likelihood estimators are often not very robust, which makes them sensitive to few particular observations, which constitutes a serious problem even in extreme value statistics. Using maximum likelihood estimator point of view, the assumption that for a Pareto-type distribution, above a certain threshold, the relative excesses behave as ordered data from a strict Pareto distribution is sometimes over-optimistic. This mostly happens when the slowly varying part disappears at a very slow rate in many instances resulting in severe bias.

It is known, that formal heavy-tailed propositions can only be satisfactorily involved for empirical constructs if sample data can be taken as a reasonable representation of the underlying distribution. In practice, distribution data may be contaminated by errors. The point of departure is recent research which has shown that Hill estimator is nonrobust. This means that small amounts of data contamination in the wrong place can reverse unambiguous conclusions. The "wrong place" usually means in the upper tail of distribution. As shown in (Brazauskas and Serfling (2000A)), small errors in the estimation of the tail index can bring large errors in the estimation of quantiles. Robust methods for extreme values have been recently addressed by literature. (Brazauskas and Serfling (2000B)) consider robust estimation in the strict Pareto model. (Vandewalle et al. (2007)) proposed robust tail index estimation procedure for the semi-parametric setting of Pareto-type distributions. As discussed in the paper (Stehlík et al.(2008)), t-estimation is at least competitive estimation technique at presence of heavy tails. In (Fabián and Stehlík (2008)) we have shown that t-estimation is clearly better when contamination is present. In this paper we study the generalization of Hill estimator based on t-estimator for Pareto and prove it to be more robust than the classical one. The main novelty of this approach is distributional sensitivity of the estimator: despite all classical modification of Hill estimator for Pareto regularly varying tails are based on asymptotics $x \to \infty$, our method is more sensitive to the interior of the distribution and thus to the distribution itself. In the recent literature there were some works on robustification of Hill estimator, however, our main aim in this paper is to construct the t-Hill like estimator, which is distributional sensitive. However, as it can be seen from this paper, as side effect we get also robustness. The proofs and technicalities are put into Appendix to maintain the better discussion.

2 Theory

It was shown in Fabián (2008) that regular continuous distributions with interval support $\mathcal{X} \in \mathbb{R}$ can be characterized, besides the cumulative distribution function $F(x)$ and probability density $f(x)$, by its t-score, given by

$$T(x) = \frac{1}{f(x)} \frac{d}{dx} \left(-\frac{1}{\eta'(x)} f(x) \right), \tag{2}$$

where $\eta : \mathcal{X} \to \mathbb{R}$ is an appropriate, strictly increasing continuous mapping. In the case of support $\mathcal{X} = (a, \infty)$, mapping

$$\eta(x) = \log(x - a) \tag{3}$$

yields often the simplest formulas for t-scores. The t-score is a suitable function for using the generalized moment method for estimation of parameters of heavy-tailed distributions, since it appeared that T is for these distributions bounded, and the moments

$$ET^k = \int_{\mathcal{X}} T(x)^k \, dF(x), \qquad k = 1, 2, ..., \tag{4}$$

exist and are often given by simple expressions. Let us call them the *t-score moments.* Particularly,

$$ET = 0 \tag{5}$$

and ET^2 is the Fisher information for x^*, which is the solution of equation

$$x^* : \qquad T(x) = 0,$$

called the *t-mean*, which can be considered as a measure of central tendency of distributions (Fabián, 2004, 2008).

Let $\theta \in \Theta \subseteq \mathbb{R}_m$ and $(X_1, ..., X_n)$ be iid sample from F_θ. The parametric version of (4) yields the generalized moment estimation equations for θ in the form

$$\hat{\theta}_n : \qquad \frac{1}{n}\sum_{i=1}^{n} T(x_i;\theta)^k = ET^k(\theta), \quad 1 \le k \le m. \tag{6}$$

Since $\hat{\theta}_n$ is the M-estimate, it is strongly consistent and asymptotically normal with the asymptotic variance-covariance matrix derived by Fabián (2001). Since distributions with heavy tails have bounded t-scores, $\hat{\theta}_n$ of heavy-tailed distributions are robust with respect to large values in observed samples.

Let us consider Pareto distribution $P(1/\alpha)$ with support $\mathcal{X} = [1, \infty)$ and density

$$f(x) = \frac{\alpha}{x^{\alpha+1}}.$$

Using the mapping $\eta = \log(x-1)$, $\eta'(x) = 1/(x-1)$ and, by (2), the t-score (2) is

$$T_\alpha(x) = -1 - (x-a)f'(x)/f(x) = \alpha(1 - x^*/x)$$

where $x^* = (\alpha+1)/\alpha$. It follows from (6) and (5) that

$$\sum_{i=1}^{n} T(x_i;\alpha) = 0$$

so that $\hat{x}^* = \bar{x}_H$ where $\bar{x}_H = n/\sum_1^n 1/x_i$ is the harmonic mean, and

$$\hat{\alpha} = 1/(\hat{x}^* - 1).$$

It suggests to introduce a variant of the Hill estimator as

$$\hat{\gamma}_k = \frac{1}{\hat{\alpha}_k} = H^*_{k,n} = \frac{1}{\frac{1}{k}\sum\limits_{j=1}^{k} \frac{X_{n-k,n}}{X_{n-j+1,n}}} - 1, \tag{7}$$

where harmonic mean is taken from the last k observed values with threshold $X_{n-k,n}$.

In the following theorem we provide the consistency of the t-Hill estimator for Pareto distribution. For proof see Appendix.

Theorem 1 *T-Hill estimator for Pareto distribution is consistent.*

3 Comparisons

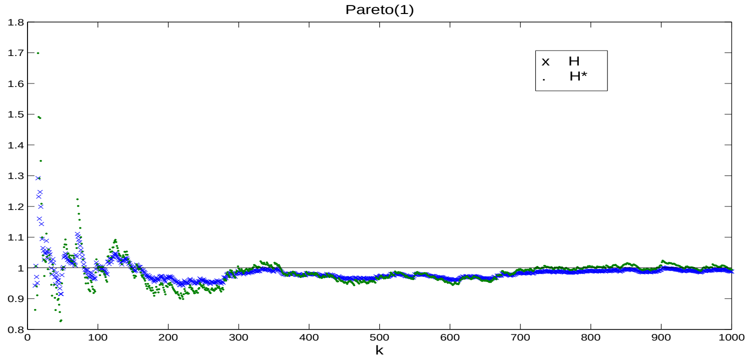

Figure 1:

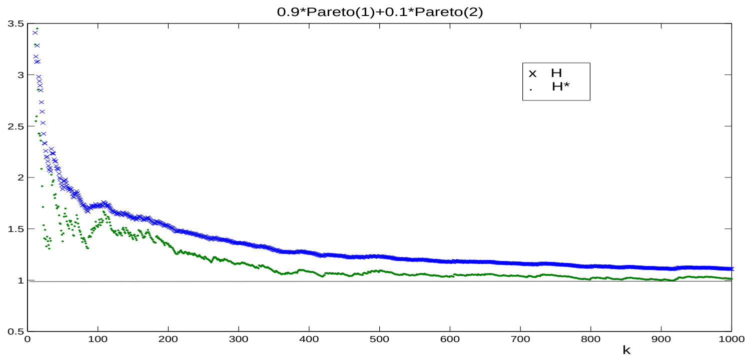

Figure 2:

One of problems with the Hill estimator is that it is not sufficiently robust. On the other hand, since the t-Hill estimator is based on harmonic mean, it is resistant to large observations so that it yields more realistic values for large k. Hill and t-Hill plots for random sample from Pareto $P(1)$ distribution are shown in Figure 1. The length of the sample was 1001 points. It is apparent that t-moment Hill estimator in his first part too much oscillates. The reason is that it is very sensitive to an abrupt change of the threshold value.

Figure 2 and Figure 3 shows Hill plot $H = \{k, H_{n,k}\}$ and t-Hill plot $H^* = \{k, H^*_{n,k}\}$ for samples generated from the contaminated Pareto distribution

$$F_c = 0.9 * P(1) + 0.1 * P(\delta)$$

with $\delta = 3$ (see Figure 2) and $\delta = 5$ (see Figure 3).

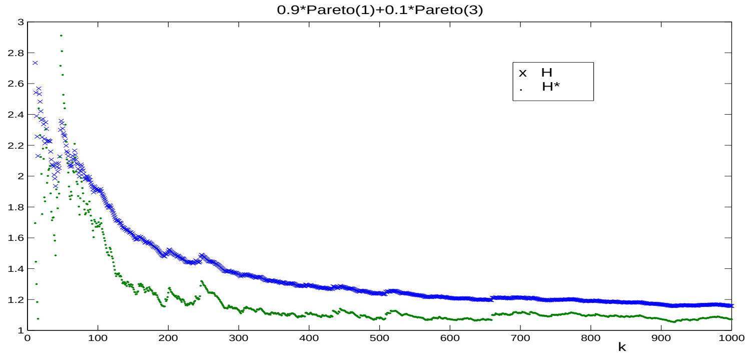

Figure 3:

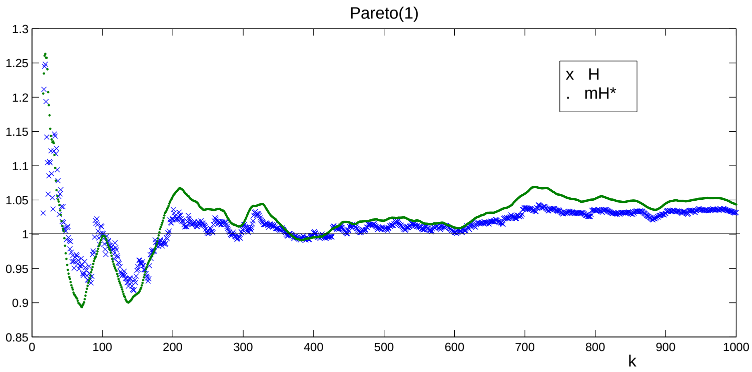

Figure 4:

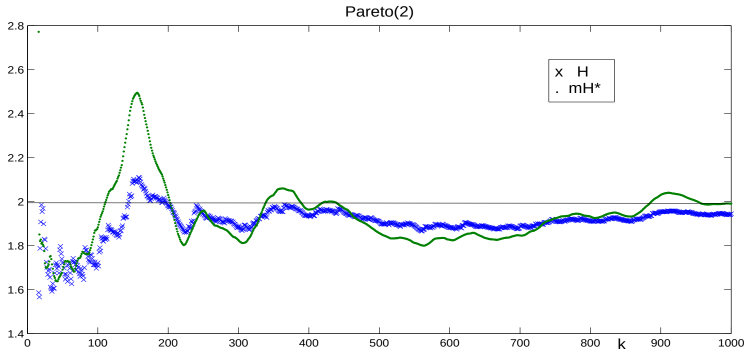

Figure 5:

It is apparent that values of t-Hill plots for large k are not too much influenced by large observed values as in ordinary Hill plots.

Oscillations of t-Hill plots can be suppressed by smoothing. Figures 4 and 5 show smoothed versions of mH^* computed simply as

$$mH_{k,n} = \frac{1}{2r+1} \sum_{j=k+1}^{k+r} H_{j-[r/2]],n}$$

(with omitting in the figures first $[r/2]$ values).

Consider now data generated from a distribution different from the Pareto one. As an example, let us consider the log-gamma distribution $L(c,\alpha)$ with support

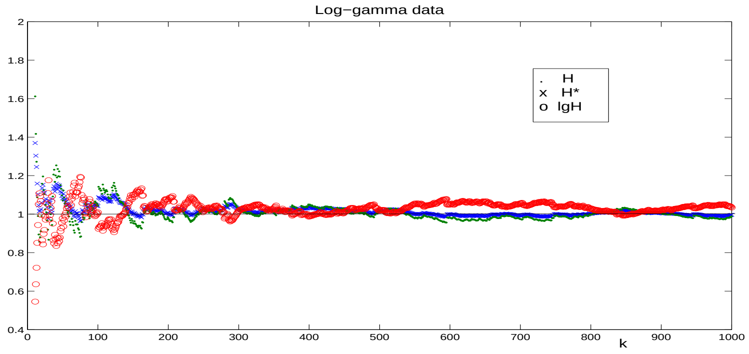

Figure 6:

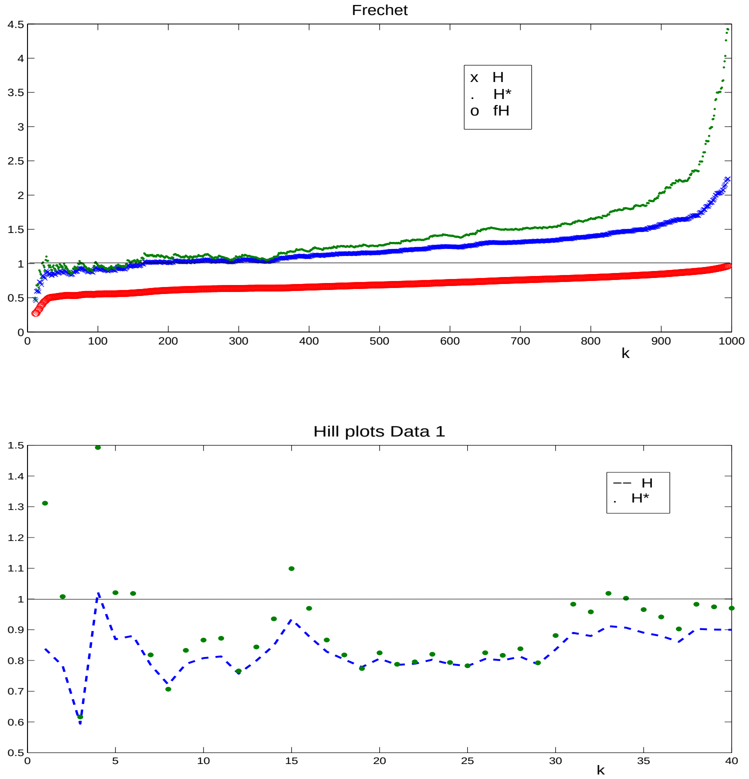

Figure 8:

$\mathcal{X} = (1, \infty)$ and density

$$f(z) = \frac{c^\alpha}{\Gamma(\alpha)} (\log z)^{\alpha-1} z^{-(c+1)}. \tag{8}$$

In this case, more simple formulas are obtained by the use of $\eta : (1, \infty) \to \mathbb{R}$ in the form

$$\eta(x) = \log(\log x).$$

Since $\eta'(x) = 1/(x \log x)$, by (2)

$$T(x) = \frac{1}{f(x)} \frac{d}{dx} \left(-(\log x)^c \alpha x^{-c} \right) = c \log x - \alpha$$

so that the 'loglog' t-mean is $x^* = e^{\alpha/c}$. As the 'second log-log moment' $ET^2 = E[c^2 \log^2(x/x^*)] = \alpha$, the estimation equations (6) are

$$\sum_{i=1}^{n} c \log x_i - \alpha = 0 \tag{9}$$

$$\sum_{i=1}^{n} (c \log x_i - \alpha)^2 = \alpha \tag{10}$$

By setting $s_1 = \frac{1}{k} \sum_{i=1}^{k} \log x_i$ and $s_2 = \frac{1}{k} \sum_{i=1}^{k} \log^2 x_i$, it follows from (9) $\hat{\alpha} = s_1 \hat{c}$ and from (10) $\hat{c}(s_2 - s_1^2) = s_1$ so that the Hill-like estimate of the tail index (cf. (Beirlant et al.(2005))) is given by closed-form expression

$$\hat{\gamma}_k = \frac{1}{\hat{c}_k} = \frac{s_2}{s_1^2} - 1.$$

The Hill-like estimates based on log-gamma distribution in Figure 6 are denoted by lgH. It appeared that for data generated from $L(1,1)$ distribution, the lgH plot gives similar results as the Hill and t-Hill plots. On the other hand, for data taken from the Fréchet distribution (Figure 7), the Fréchet Hill-like estimator underestimates the true value $\gamma = 1$, whereas the last part of both H and H^* it overestimate.

Real data illustrative example deals with data for Example 1, (Stehlík et al.(2008)). These data consist of 96 payments in one year in non life insurance. At Figure 8 the estimation for this data is illustrated.

4 Conclusions

There are two main ways of avoiding misleading conclusions due to nonrobust tools in the presence of contaminated data. One is based on statistics that automatically remove from the sample data that are potentially troublesome. The other relies on the specification of parametric models for the distribution of the data and uses robust estimators of the parameters. As can bee seen from this letter, t-Hill estimator of Pareto tail index is distribution sensitive and "naturally" robust. If more accurate fit to the central part of distribution is needed, we suggest to use e.g. combining a Pareto estimate of the upper tail with a non-parametric estimate of the rest of the distribution, as suggested by (Cowell and Victoria-Feser (2007)) and by (Davidson and Flachaire (2007)) with bootstrap methods. The derivation of distribution of t-Hill estimator, comparison of its efficiency with other estimators together with its comparison to a different robust estimators will be worth further investigation.

5 Appendix

Proof of Theorem 1

Consistency of the tail empirical measure, defined as a random element of $M_+(0,\infty]$, the space of nonnegative Radon measures on $(0,\infty]$, implies the consistency of t-Hill estimator for $1/(\alpha+1)$. The proof proceeds by a series of steps following the proof of classical Hill estimator in (Resnick (2007)).

STEP 1): consistency of the empirical measure (given in 4.14 by (Resnick (2007))) implies $\frac{X(k)}{b(\frac{n}{k})} \xrightarrow{P} 1$ as $n \to \infty, k \to \infty, \frac{k}{n} \to 0$.

This allows us to consider $X(h)$, as a consistent estimator of $b(\frac{n}{h})$.

STEP 2): In $M_+(0,\infty] : v_n \xrightarrow{P} v_\alpha$ as $n \to \infty, k \to \infty, \frac{h}{n} \to 0$. This is proved by a scaling argument.

Define the operator $T : M_+(0,\infty] \times (0,\infty) \to M_+(0,\infty]$ by $T(\mu,x)(A) = \mu(xA)$. From (4.14) in (Resnick (2007)) and Proposition 3.1 therein we get joint weak convergence $(v_n, \frac{X(k)}{b(\frac{n}{k})}) \Longrightarrow (v_\alpha, 1)$ *in* $M_+(0,\infty) \times (0,\infty)$.

Since $v_n^1(.) = T(v_n, \frac{X(k)}{b(\frac{n}{k}})$, the conclusion will follow by the continuous mapping theorem and continuity of the operator T at $(v_\alpha, 1)$.

STEP 3): Integrate the tails of the measure against $x^{-2}dx$. The integral functional is continuous on $[1, M]$, for any M and so it is only on $[M,\infty]$ that care must be exercised. By the 2nd converging Theorem, only we must show that

$$\lim_{M\to\infty} \lim_{n\to\infty} P\left[\int_M^\infty v_n^1(x,\infty)x^{-2}dx > \delta\right] = 0$$

We have

$$P\left[\int_M^\infty v_n^1(x,\infty)x^{-2}dx > \delta\right] \le I + II$$

where $II \le P\left[|\frac{\hat{b}(n/k)}{b(n/k)} - 1| \ge \eta\right] \to 0$ and

$I \le P\left[\int_M^\infty v_n((1-\eta)x,\infty)x^{-2}dx > \delta\right] = P\left[\int_{M(1-\eta)}^\infty v_n(x,\infty)x^{-2}dx > \delta\right]$

and this probability has a bound from Markov's inequality $\delta^{-1}E\left[\int_{M(1-\eta)}^\infty v_n(x,\infty)x^{-2}dx\right] = \delta^{-1}\int_{M(1-\eta)}^\infty \frac{n}{k}P(X_1 > b(n/k)x)x^{-2}dx \to \delta^{-1}\int_{M(1-\eta)}^\infty x^{-2-\alpha}dx$ for $n \to \infty$.

Finally $\delta^{-1}\int_{M(1-\eta)}^\infty x^{-2-\alpha}dx = \frac{const}{M^{\alpha+1}} \to 0$ for $M \to \infty$.

We have applied Karamata's theorem (see Thereom 2.1 in (Resnick (2007)), page 25.).

STEP 4): We have proved that $\int_1^\infty v_n^1(x,\infty]x^{-2}dx \xrightarrow{P} \int_1^\infty v_\alpha(x,\infty]x^{-2}dx$.
So $\int_1^\infty v_n^1(x,\infty]x^{-2}dx$ is a consistent estimator of $\frac{1}{\alpha+1}$ and we just need to see that this is indeed the modified Hill estimator. This is done as follows:

$$\int_1^\infty v_n^1(x,\infty]x^{-2}dx = \int_1^\infty \frac{1}{h}\sum_{i=1}^n \varepsilon(x_i/b^1(n,k))(x,\infty]x^{-2}dx =$$

$$= \frac{1}{k}\sum_{i=1}^n \int_1^{X_i/b(n,k)v1} x^{-2}dx = 1 - \frac{1}{h}\sum_{i=1}^n \frac{1}{\frac{X(i)}{X(h)}}$$

Acknowledgement

Research was supported by project Aktion No 50p14.

References

Brazauskas V. and Serfling R. Robust and efficient estimation of the tail index of a single-parameter Pareto distribution. North Amer. Actuar. J. 4, 12-27.

Brazauskas V. and Serfling R. Robust estimation of tail parameters for two-parameter Pareto and exponential models via generalized quantile statistics. Extremes 3:3, 231-249.

Cowell F.A. and Victoria-Feser, M.P. Robust Lorenz Curves: A Semiparametric Approach, Journal of Economic inequality, 5, 21-35

Davidson R. and Flachaire E. (2007) Asymptotic and bootstrap inference for inequality and poverty measures, Journal of Econometrics 141 (2007) 141–166

Fabián Z. Induced cores and their use in robust parametric estimation, *Communication in Statistics, Theory Methods*, **30** (2001), pp.537-556.

Fabián Z. Estimation of simple characteristics of samples from skewed and heavy-tailed distribution, in *Recent Advances in Stochastic Modeling and Data Analysis* (ed. Skiadas, C.), Singapore, World Scientific (2007), pp.43-50.

Fabián Z. New measures of central tendency and variability of continuous distributions, *Communication in Statistics, Theory Methods* **37** (2008), pp.159-174.

Fabián Z. Parametric estimation by generalized moment method for extremes, Communications in Dependability and Quality Management 11(2008)4 , 26-35.

Fabián Z. and Stehlík M. (2008), A note on favorable estimation when data is contaminated, Communications in Dependability and Quality Management 11(2008)4, 36-43.

Stehlík M., Potocký R., Waldl H. and Fabián Z. (2008) On the favourable estimation of fitting heavy tailed data, IFAS res. report Nr. 32

Reiss R.D. and Thomas, M. Statistical analysis of extreme values, Birkhauser verlag, Berlin

Beirlant J., Goegebeur Y., Segers J. and Teugels, J. (2005). Statistics of Extremes. Theory and applications. Wiley.

Resnick S.I. Heavy-tail phenomena (2007). Springer.

Vandewalle B., Beirlant J., Christmann A. and Hubert M., A robust estimator for the tail index of Pareto-type distributions, Computational Statistics & Data Analysis 51, 6252-6268.