On the tail index inference based on the scaling function method

Danijel Grahovac[*1], **Mofei Jia**[2], **Nikolai Leonenko**[3] **and Emanuele Taufer**[2]

[1] *Department of Mathematics, Josip Juraj Strossmayer University, Osijek, Croatia*
[2] *Department of Economics and Management, University of Trento, Italy*
[3] *Cardiff School of Mathematics, Cardiff University, UK*

Abstract

In [3], a new method has been presented for making inference about the tail of samples coming from unknown heavy-tailed distribution. Method is based on asymptotic properties of the empirical structure function, a variant of statistic that resembles usual sample moments. Using this approach one can successfully inspect the nature of the tail of the underlying distribution, as well as provide estimated values on the unknown tail index. Here we briefly describe the method and test its performance on some simulated and real world data by comparing it with the well known Hill estimator.

Keywords: heavy-tailed distributions, tail index, empirical structure function, scaling functions, Hill estimator.
AMS subject classifications: 62F10, 62F12, 62E20.

1 Introduction

Heavy-tailed distributions are of considerable importance in modeling a wide range of phenomena in finance, geology, hydrology, physics, queuing theory and telecommunication. Since the work of Mandelbrot [4], where stable distributions with index less than 2 have been advocated for describing fluctuations of cotton prices, there has been an exhausting research concerning the use of heavy-tailed distribution in the context of finance.

We define that the distribution of some random variable X is heavy-tailed with index $\alpha > 0$ if it has a regularly varying tail with index $-\alpha$, i.e.

$$P(|X| > x) = \frac{L(x)}{x^\alpha}, \quad |x| \to \infty,$$

where $L(t), t > 0$, is a slowly varying function, i.e., $L(tx)/L(x) \to 1$ as $|x| \to \infty$, for every $t > 0$. In particular, this implies that $E|X|^q < \infty$ for $q < \alpha$ and $E|X|^q = \infty$ for $q > \alpha$, which can be used as the alternative definition. We are interested in the estimation of the unknown tail index α, measuring the "thickness" of the tails, based on the finite data sample with no additional assumptions on the distribution of the data.

There exists a range of estimators for this particular problem. The most well known estimators are the Pickand's, Hill and moment estimator by Dekkers, Einmahl and de Haan. A nice survey of these estimators and their properties can be found in [2] and [1]. Tail index estimators are usually based on upper order statistics and their asymptotic properties. As an alternative, [5] proposed an estimator based on the asymptotics of the partial sum. In this paper we present a novel approach given in [3]. We evaluate the performance of this estimator on some real world examples and compare it with probably the most popular one, the Hill estimator.

*Corresponding author, e-mail: dgrahova@mathos.hr

2 Estimation method

The estimator presented in [3] is based on asymptotic properties of the empirical structure function (also called partition function), a kind of statistic that resembles usual sample moments. More precisely, given a sample $X_1, \ldots, X_n$ coming from a strictly stationary stochastic process $\{X_t, t \in \mathbb{Z}_+\}$ (discrete time) or $\{X_t, t \in \mathbb{R}_+\}$ (continuous time) which has a heavy-tailed marginal distribution with unknown tail index α, define

$$S_q(n,t) = \frac{1}{\lfloor n/t \rfloor} \sum_{i=1}^{\lfloor n/t \rfloor} \left| \sum_{j=1}^{\lfloor t \rfloor} X_{t(i-1)+j} \right|^q, \tag{1}$$

where $q > 0$ and $1 \leq t \leq n$. In words, we partition the data into consecutive blocks of length $\lfloor t \rfloor$, then sum each block and take the power q of the absolute value of the sum. Finally, we average over all $\lfloor n/t \rfloor$ blocks. Notice that for $t = 1$ one gets the usual empirical q-th absolute moment.
Asymptotic properties of $S_q(n,t)$ have been considered before in the context of multifractality detection (see [3] and the references therein). Instead of keeping t fixed, we take it to be of the form $t = n^s$ for some $s \in (0,1)$, which allows the blocks to grow as the sample size increases. It is clear that then $S_q(n, n^s)$ will diverge since $s > 0$. The quantity of interest is the rate of divergence of this statistic, i.e. we consider the limiting behavior of $\ln S_q(n, n^s)/\ln n$. This has been established in [3] under the assumptions of strict stationarity of the sequence $X_t, t \in \mathbb{Z}_+$ and mild dependence condition in the form of the strong mixing property with an exponentially decaying rate (for details see [3]). It is also assumed that the expectation is zero in case when it is finite. The proof of the theorem can be found in [3].

Theorem 2.1. *Suppose $X_t, t \in \mathbb{Z}_+$ is a strictly stationary sequence that has a strong mixing property with an exponentially decaying rate and suppose that $X_t, t \in \mathbb{Z}_+$ has a heavy-tailed marginal distribution with tail index $\alpha > 0$. Suppose also that $EX_i = 0$ when $\alpha > 1$. Then for $q > 0$ and every $s \in (0,1)$*

$$\frac{\ln S_q(n, n^s)}{\ln n} \xrightarrow{P} R_\alpha(q,s) := \begin{cases} \frac{sq}{\alpha}, & \text{if } q \leq \alpha \text{ and } \alpha \leq 2, \\ s + \frac{q}{\alpha} - 1, & \text{if } q > \alpha \text{ and } \alpha \leq 2, \\ \frac{sq}{2}, & \text{if } q \leq \alpha \text{ and } \alpha > 2, \\ \max\left\{ s + \frac{q}{\alpha} - 1, \frac{sq}{2} \right\}, & \text{if } q > \alpha \text{ and } \alpha > 2, \end{cases} \tag{2}$$

as $n \to \infty$, where $\xrightarrow{P}$ stands for convergence in probability.

It is clear that the limit considered in the preceding theorem heavily depends on the tail index α, which makes it possible to make inference about the unknown tail index. First notice that if for some non-negative sequence $\{Z_n\}$ of random variables $\ln Z_n/\ln n \xrightarrow{P} a \in \mathbb{R}$, then for some function M such that $\ln M(n)/\ln n \to 0$, $Z_n/n^a M(n) \xrightarrow{d} Z$ as $n \to \infty$, where Z is a random variable not identically equal to zero (possible degenerate). So, it follows from Theorem 2.1 that $\varepsilon_n := \frac{S_q(n,n^s)}{n^{R_\alpha(q,s)} M(n)} \xrightarrow{d} \varepsilon$, where ε is a random variable not identically equal to zero. By simply rewriting this, one arrives at the

$$\frac{\ln S_q(n, n^s)}{\ln n} = R_\alpha(q,s) + \frac{\ln M(n)}{\ln n} + \frac{\ln \varepsilon_n}{\ln n}. \tag{3}$$

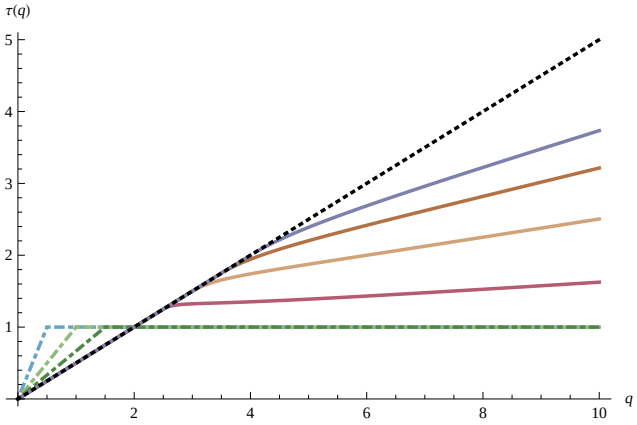
This equation can be seen as the regression model. The term $\ln \varepsilon_n/\ln n$ can be considered as an error term in the regression of $\ln S_q(n, n^s)/\ln n$ on q and s. One should count on the intercept in the model, in order to compensate for the $\ln M(n)/\ln n$ term. The possible nonzero mean of an error can be subtracted and considered as a part of the intercept.
The basic idea of the approach presented in [3] is to estimate the tail index α by the means of Equation (3). To avoid bivariate regression, one can assume the limit is linear in s, i.e. $R_\alpha(q,s) = \tau(q)s + c(q)$. This

holds exactly except in the case $q > \alpha > 2$. By theoretically regressing $\ln S_q(n, n^s)/\ln n$ on s, for a range of values $s \in (0, 1)$, one gets the expression for $\tau(q)$ (notice that this is obvious in case $\alpha \leq 2$):

$$\tau(q) = \begin{cases} \frac{q}{\alpha}, & \text{if } 0 < q \leq \alpha \ \& \ \alpha \leq 2, \\ 1, & \text{if } q > \alpha \ \& \ \alpha \leq 2, \\ \frac{q}{2}, & \text{if } 0 < q \leq \alpha \ \& \ \alpha > 2, \\ \frac{q}{2} + \frac{2(\alpha-q)^2(2\alpha+4q-3\alpha q)}{\alpha^3(2-q)^2}, & \text{if } q > \alpha \ \& \ \alpha > 2. \end{cases} \tag{4}$$

$\tau(q)$ is refereed to as the scaling function. When α is large, i.e., $\alpha \to \infty$, it follows from (4) that $\tau(q) = q/2$. This corresponds to data coming from a distribution with all moments finite, e.g., an independent normally distributed sample. This line will be referred to as the baseline. Theoretical plots of scaling functions for a range of α values are shown in Fig. 1. It is clear that the shape of the scaling function is heavily influenced by the value of tail index α.

Figure 1: Plots of scaling function $\tau(q)$ against the moment q



The baseline is shown by a dashed line. The case $\alpha \leq 2$ ($\alpha = 0.5, 1.0, 1.5$) and $\alpha > 2$ ($\alpha = 2.5, 3.0, 3.5, 4.0$) are shown by dot-dashed and solid lines, respectively.

Having a finite data sample, one can estimate $\tau(q)$ in a single point q as the slope in the simple linear regression model by regressing $\ln S_q(n, n^s)/\ln n$ on s, for a range of values of $s \in (0, 1)$. More precisely, fix $q > 0$ and for $s_i \in (0, 1),\ i = 1, \ldots, m$ calculate $S_i = \ln S_q(n, n^{s_i})/\ln n,\ i = 1, \ldots, m$ based on the data sample. Now, estimate the value of the scaling function at the point q as

$$\left(\hat{\tau}(q), \hat{b}\right) = \underset{(a,b)\in\mathbb{R}^2}{\arg\min} \sum_{i=1}^{m} \left(S_i - a s_i - b\right)^2. \tag{5}$$

Repeating this for a range of q makes it possible to give a plot of empirical scaling function $\hat{\tau}$. By comparing empirical scaling function with Fig. 1, one can make inference about the nature of the tails of the underlying distribution. Moreover, by minimizing the difference between the theoretical scaling function (4) and the empirical one $\hat{\tau}(q)$ for some range of $q \in (0, q_{max})$ one can find the estimate for α. More precisely, for points $q_i \in (0, q_{max}),\ i = 1, \ldots, n$, estimate $\tau_i = \hat{\tau}(q_i)$ by the means of Equation (5). Estimator is defined as

$$\hat{\alpha} = \underset{\alpha\in(0,\infty)}{\arg\min} \sum_{i=1}^{m} \sum_{j=1}^{k} (\tau_i - \tau(q_i))^2. \tag{6}$$

Method is divided in two cases, $\alpha \leq 2$ and $\alpha > 2$, in order to simplify the estimation procedure. Cases can be distinguished graphically by plotting the empirical scaling function.

3 Examples and comparison with Hill estimator

In this section we test the performance of estimator (6) on some known data sets and compare it with the Hill estimator. Let $X_{(1)} \geq X_{(2)} \geq \cdots \geq X_{(n)}$ denote the order statistics of the sample $X_1, X_2, \cdots X_n$, and k_n be a sequence of positive integers satisfying $1 \leq k_n < n$, $\lim_{n\to\infty} k_n = \infty$, and $\lim_{n\to\infty} (k_n/n) = 0$. The Hill estimator based on k_n upper order statistics is

$$\hat{\alpha}_{k_n} = \left(\frac{1}{k_n} \sum_{i=1}^{k_n} \log \frac{X_{(i)}}{X_{(k_n+1)}} \right)^{-1}. \tag{7}$$

Hill estimator is known to be weakly consistent as well as strongly consistent and asymptotically normal under certain conditions. For details see [2]. However, performance of the Hill estimator is heavily influenced by the choice of k_n. There is no generally accepted method on how to choose k_n, and it is usually recommended to plot the values for a range of k_n values and to look for the part of the graph where the value stabilizes. The resulting plot is usually called Hill plot.

3.1 Example 1 - non-constant slowly varying function in the tail

Hill estimator is known to behave poorly if the slowly varying function in the tail is far away from constant. We compare this behavior with the performance of the estimator (6). Consider two distribution F_1, F_2 defined by their survival functions

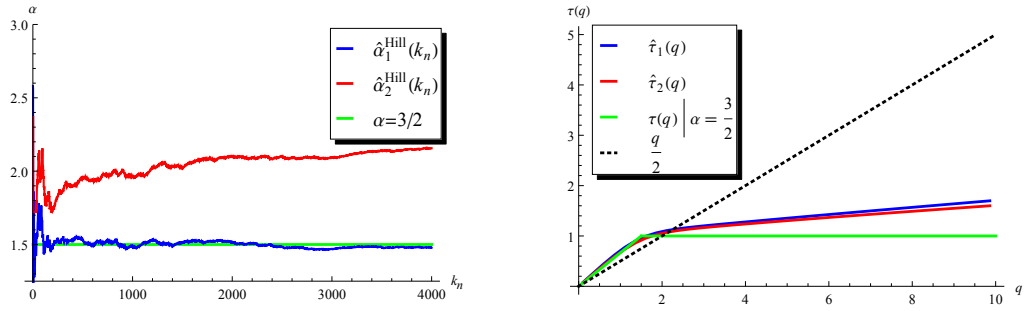
$$\overline{F}_1(x) = 1 - F_1(x) = \frac{1}{x^{\frac{3}{2}}}, \quad x \geq 1, \tag{8}$$

$$\overline{F}_2(x) = 1 - F_2(x) = \frac{e^{\frac{3}{2}}}{x^{\frac{3}{2}} \ln x}, \quad x \geq e. \tag{9}$$

Both distributions are heavy-tailed with tail index equal to $3/2$. We generate samples from these two distributions with 5000 observations. Corresponding Hill plots are shown in Figure 2(a). For F_2, one could wrongly conclude that the value of the tail index is around 2. The Hill's method is highly sensitive to the presence of non-constant slowly varying function in the tail. This is sometimes called Hill horror plot (see [2]). Figure 2(b) shows empirical scaling functions for the same samples together with the theoretical one and the baseline. One can see that scaling functions almost coincide with the theoretical one. Calculating estimates using (6) yields values $\hat{\alpha}_1 = 1.441$ and $\hat{\alpha}_2 = 1.5141$. It seems that non-constant slowly varying function affects the estimation but the effect is not so dramatical as for the Hill estimator. Most important part of the scaling functions for the inference about the tail is before the breakpoint and the breakpoint itself. For example, one can try estimating α only based on the values of $\hat{\tau}(q)$ for q less than a breakpoint observed graphically by fitting simple linear regression through origin. Theoretically, slope of the regression line should be $1/\alpha$. For example above, using $q \in (0, 1.5)$ one gets estimates for α: 1.454 for F_1 and 1.527 for F_2.

3.2 Example 2 - non heavy-tailed distribution

For the next example we compare the behavior of two estimators when the underlying distribution is not heavy-tailed. For this purpose, sample of 2000 observations was generated from standard logistic distribution



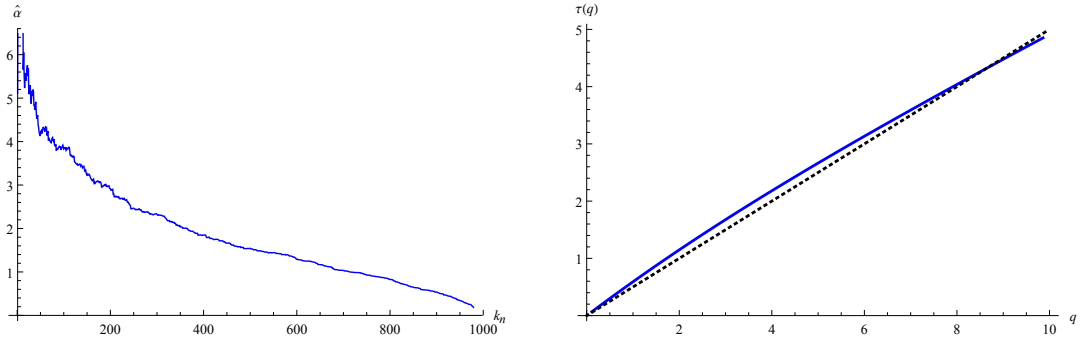
(a) Hill plot (b) Scaling functions

Figure 2: Example 1

given by probability density function

$$f(x) = \frac{e^{-x}}{(1+e^{-x})^2},\ x \in \mathbb{R}.$$

Figure 3(a) shows the Hill plot. It is impossible to draw any conclusion by only analyzing the Hill plot. This is why it is always necessary to use some other techniques for detecting heavy tails in data samples. On the other hand, estimated scaling function provides self contained characterization of the tail. From Figure 3(b) one can surely doubt the existence of heavy-tails since the empirical scaling function almost coincides with the baseline $q/2$.

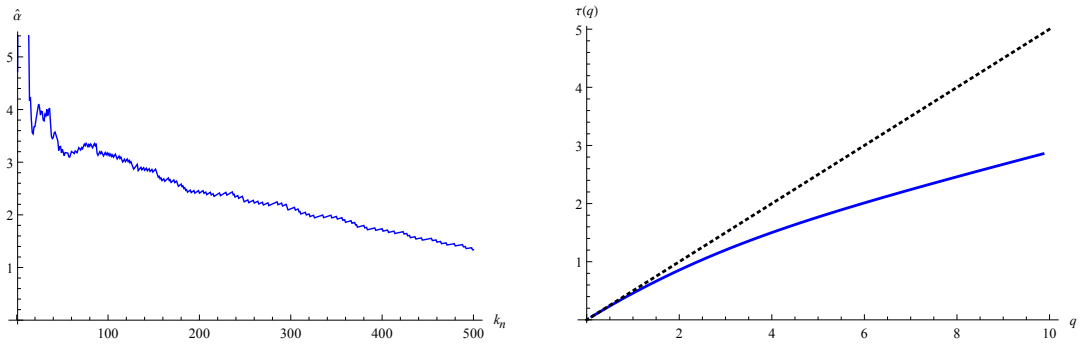


(a) Hill plot (b) Scaling function

Figure 3: Scaling function

3.3 Example 3 - EUR/USD exchange rates

In this example we analyze daily closing rates of euro against U.S. dollar during the period $2007 - 2012$. The data consists of differences of rates and has 1868 observations. Hill plot is shown in Figure 4(a) and corresponding scaling function in the Figure 4(b). Hill plot fails to stabilize, but one could say this happens for k_n around 100 yielding, for example, value $\hat{\alpha} = 3.133$ for $k_n = 100$. Scaling function evidently points that the variance is finite since the break occurs after $q = 2$ and the plot coincides with the baseline before the break. Estimation for the case $\alpha > 2$ yields the value 3.112, consistent with the Hill estimator.

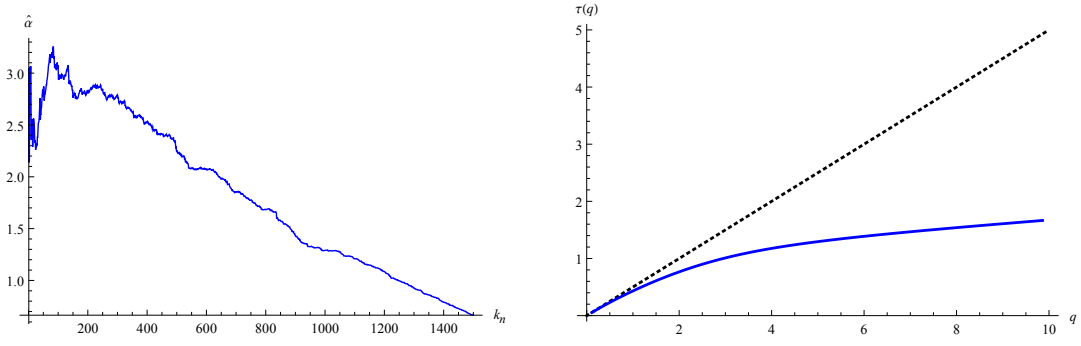


(a) Hill plot (b) Scaling functions

Figure 4: Scaling function

3.4 Example 4 - daily log-returns of DAX

Next example again involves financial data. We use daily log-returns of the German stock index DAX (September 20, 1988 - August 24, 1995), similar to Figure 6.4.12 in [2]. Hill plot in Figure 5(a) is made by using absolute value of the data. Following [2], one can conclude that the plot stabilizes around 2.8 for $100 \leq k_n \leq 300$. However, plot fails to stabilize for larger k_n, similar as in the Example 1. Data has been centered for the estimation of the scaling function on Figure 5(b). Plot shows that α could be somewhere between 2 and 2.5. Calculating the estimate (6) yields the value 2.465. Thus, there is a significant discrepancy between two estimates. Considering the inconclusiveness of the Hill plot, one could give preference to the estimate (6).



(a) Hill plot (b) Scaling function

Figure 5: Scaling function

Bibliography

[1] L. De Haan and A. Ferreira. *Extreme Value Theory: An Introduction*. Springer, 2006.

[2] Paul Embrechts, Claudia Klüppelberg, and Thomas Mikosch. *Modelling extremal events: for insurance and finance*, Volume 33. Springer Verlag, 1997.

[3] Danijel Grahovac, Mofei Jia, Nikolai N Leonenko, and Emanuele Taufer. Asymptotic properties of the partition function and applications in tail index inference of heavy-tailed data. *arXiv preprint arXiv:1310.0333*, 2013.

[4] B. Mandelbrot. The variation of certain speculative prices. *Journal of Business*, 36(4):394–419, 1963.
[5] M.M. Meerschaert and H.P. Scheffler. A simple robust estimation method for the thickness of heavy tails. *Journal of Statistical Planning and Inference*, 71(1):19–34, 1998.