On a new class of the tail index estimators addapted to the high frequency financial data

Paulauskas Vygantas
Vilnius university, Department of Mathematics and Informatics,
Naugarduko 24,
Vilnius 03225, Lithuania
E-mail: vygantas.paulauskas@mif.vu.lt

The paper presents the extended version of the talk to be presented on the Special Topics Session "Statistical analysis of extremes and its application" of the 58-th World Statistics Congress of ISI and is based mainly on the papers [10], [11].

We consider the tail index estimation, the problem which during last several decades attracted attention of theoretical statisticians and practitioners as well since in many fields of applied probability the so-called heavy-tailed distributions play an important role. Indirectly this can be confirmed by Google Scholar giving approximately 170000 entries for "tail index estimation". The problem (in its the most simple form) can be formulated as follows. Let us consider a sample $X_1, \dots, X_N$ of size N taken from a heavy-tailed distribution function (d.f.) F, that is, we assume that $X_1, \dots, X_N$ are independent identically distributed (i.i.d.) random variables with a d.f. F satisfying the following relation for large x:

$$(1) \qquad 1 - F(x) = x^{-\alpha} L(x).$$

Here $\alpha > 0$, $L(x) > 0$ for all $x > 0$ and$L$ is a slowly varying at infinity function:

$$\lim_{x \to \infty} \frac{L(tx)}{L(x)} = 1.$$

The problem is to estimate the parameter α. It is worth to mention that this is particular case of more general problem of estimation of the extreme value index in the Extreme Value Theory (Google Scholar gives approximately 670000 entries for "estimation of extreme value index "). We recall that a distribution function F is in domain of attraction of the extreme value distribution $G_\gamma(x) = \exp(-(1+\gamma x)^{-1/\gamma}),\ \gamma \in R$ if there exist constants $a_n > 0$ and b_n, such that

$$\lim_{n \to \infty} F^n(a_n x + b_n) = G_\gamma(x),$$

for all $1 + x\gamma > 0$. The parameter $\gamma \in R$ is called the extreme value index. Tail index estimation problem corresponds to the case $\gamma > 0$, and there is a simple relation between these two indices, namely, $\alpha = \gamma^{-1}$. Most popular estimators of γ are based on rank statistics and some facts from Extreme Value Theory, such as limit behavior of excess distribution function and the relation between order-statistics and exponential distributions (the Renyi representation theorem). For details and explanation of intuition, on which several popular estimators , such as Hill's, Pickand's and others, are based, we refer to a recent paper [15]. Some of estimators, (for example, Hill's estimator) , are designated only for tail index estimation, that is, for positive γ, others work both for positive and negative γ. But all they are based on idea that it is necessary to take the largest values from the ordered statistics $X_{N,1} \le X_{N,2} \le \cdots \le X_{N,N}$ from a sample $X_1, \dots, X_N$, and from these values to extract information about parameter γ.

From now on we concentrate on the tail index estimation only.

Our goal is to present one more estimator of the tail index and to discuss its merits and shortcomings. Although it was introduced almost ten years ago (see [1] and [9]), only recently it was realized that due to its construction this estimator can be successfully used in analysis of high frequency data, which became rather important issue in financial econometrics, network statistical analysis, and some other fields. At first we present the construction of this estimator and the intuition behind it.

We divide a sample into n groups $V_1, \dots, V_n$, each group containing m random variables, that is, we assume that $N = n \cdot m$ and $V_i = \{X_{(i-1)m+1}, \dots, X_{im+1}\}$. (In practice, at first m is chosen and then $n = [N/m]$ is taken, where $[x]$ stands for the integer part of a number $x > 0$.) Let $M_{n,i}^{(1)} = \max\{X_j \colon X_j \in V_i\}$ and let $M_{n,i}^{(2)}$ denote the second largest element in the same group V_i. Let us denote

$$(2) \qquad \upsilon_{n,i} = \frac{M_{n,i}^{(2)}}{M_{n,i}^{(1)}}, \qquad S_n = \sum_{i=0}^{n} \upsilon_{ni}, \qquad Z_n = n^{-1} S_n.$$

In [1] the estimator Z_n from (2) (in a different context of a sample from multivariate stable distribution and with the restriction $0 < \alpha < 1$) was based on the following relation (see LePage et al [**?**])

$$(M_{n,i}^{(1)}, M_{n,i}^{(2)}) m^{-1/\alpha} \xrightarrow{\mathrm{D}}_{N\to\infty} (\Gamma_1^{-1/\alpha}, \Gamma_2^{-1/\alpha}),$$

and the fact that

$$E\Big(\frac{\lambda_1}{\lambda_1 + \lambda_2}\Big)^{1/\alpha} = \frac{\alpha}{\alpha + 1}.$$

Here $\Gamma_i = \sum_{j=1}^{i} \lambda_j, \quad \lambda_j, j \geq 1$, are i.i.d. standard exponential random variables, and $\xrightarrow{\mathrm{D}}$ denotes the convergence in distribution.

In [9] it was noted that estimator from (2) can be based on a different idea. Let us take two independent random variables X and Y with the same Pareto distribution

$$F(x) = 1 - C_1 x^{-\alpha}, \quad x \geq C_1^{1/\alpha},$$

and denote

$$(3) \qquad W = \frac{\min(X, Y)}{\max(X, Y)}.$$

It is not difficult to verify that, denoting $p = \alpha/(1 + \alpha)$, we have $\mathbf{E}\, W = p$ (since W is invariant under scale transformation, we can take $C_1 = 1$ and, in the sequel, we shall refer to that case as a standard Pareto distribution). Therefore, in the case of the Pareto distribution, quantity Z_n, as an estimator for the parameter p (we shall denote this quantity by $\hat{p}$, and as in [13], we shall call it as the DPR estimator), is nothing but the sample mean for a bounded random variable, moreover, in this case the best choice is to take m=2. If the underlying distribution F is not exactly Pareto, but satisfies condition (1) then it is natural to expect that for large m $\mathbf{E}\, \hat{p} = \mathbf{E}\, \upsilon_{n1}$ will be close to p, and Z_n will be consistent estimator of this parameter. But it is clear that having only condition (1) without any additional information about the function L, it is difficult to get good properties, such as the asymptotic normality, of any estimator of the parameter α. Therefore from now we assume that a sample is taken from a distribution function F which satisfies the second-order asymptotic relation (as $x \to \infty$)

$$(4) \qquad 1 - F(x) = C_1 x^{-\alpha} + C_2 x^{-\beta} + \mathrm{o}(x^{-\beta}),$$

with some parameters $0 < \alpha < \beta \leq \infty$. (Here it is possible to mention that in most papers dealing with extreme value index estimation, a little bit general the second-order condition with a different parametrization is used).

Since Z_n is a sum of i.i.d. and bounded random variables, main difficulty in proving asymptotic normality of estimator $\hat{p}$ is to get good estimate of the bias $\gamma_m = \mathbf{E}\, \hat{p} - p$. The following estimate was obtained in [9]:

$$(5) \qquad |\gamma_m| \leq C_0 m^{-\zeta},$$

where $\zeta = (\beta - \alpha)/\alpha$, and C_0 is a constant depending on C_1, C_2, α, and β. This estimate allowed to prove the asymptotic normality of the DPR estimator. Asymptotic of γ_m was obtained in [10] and the following result was proved. We write $a_n \sim b_n$ if $\lim_{n\to\infty} a_n b_n^{-1} = 1$.

Theorem 1 [10]. *Let us suppose that F satisfies* (4) *with* $0 < \alpha < \beta < \infty$ *and* $C_1 > 0$. *Then we have*

$$\gamma_m \sim \chi m^{-\zeta}, \quad (m \to \infty), \tag{6}$$

where

$$\chi = \frac{C_2 \beta \zeta \Gamma(\zeta+1)}{C_1^{\zeta+1}(\alpha+1)(\beta+1)}.$$

For sufficiently large N (ensuring that $m \geq 2$) taking

$$m = \left(\frac{2\zeta\chi^2}{\sigma^2}\right)^{1/(1+2\zeta)} N^{1/(1+2\zeta)} \tag{7}$$

we get that MSE (mean square error) is minimal

$$\mathbf{E}\,(\hat{p} - p)^2 \sim (1+2\zeta)\left(\frac{\chi^2\sigma^{4\zeta}}{(2\zeta)^{2\zeta}N^{2\zeta}}\right)^{1/(1+2\zeta)}. \tag{8}$$

Under this choice of m we have the asymptotic normality

$$\sqrt{n}(\hat{p} - p) \xrightarrow{\mathrm{D}}_{N\to\infty} N(\mu, \sigma^2), \tag{9}$$

where

$$\mu = \sigma(2\zeta)^{-1/2}\,\mathrm{sgn}(\chi), \quad \sigma^2 = \frac{\alpha}{(\alpha+1)^2(\alpha+2)}.$$

This result allowed to compare the introduced estimator $\hat{p}$ with some other well-known estimators of the tail index. In [5] there were compared four estimators $\gamma_{N,k}^{(i)}$, $i = 1,2,3,4$, of the parameter γ:

$$\gamma_{N,k}^{(1)} = \frac{1}{k}\sum_{i=0}^{k-1} \log X_{N,N-i} - \log X_{N,N-k},$$

$$\gamma_{N,k}^{(2)} = (\log 2)^{-1} \log \frac{X_{N,N-[k/4]} - X_{N,N-[k/2]}}{X_{N,N-[k/2]} - X_{N,N-k}},$$

$$\gamma_{N,k}^{(3)} = \gamma_{N,k}^{(1)} + 1 - \tfrac{1}{2}(1 - (\gamma_{N,k}^{(1)})^2/M_N)^{-1}, \quad \gamma_{N,k}^{(4)} = \frac{M_N}{2\gamma_{N,k}^{(1)}},$$

where

$$M_N = \frac{1}{k}\sum_{i=0}^{k-1}(\log X_{N,N-i} - \log X_{N,N-k})^2.$$

Under the assumption (4) all estimators $\gamma_{N,k}^{(i)}$, $i = 1,2,3,4$, are asymptotically normal and, more important, all they have the same rate of convergence. In [5] the asymptotic mean square error (AMSE) was chosen as a criterion of comparison of estimators , and all these estimators were compared. It turned out that none of these estimators dominates the others: for different values of the parameters, present in the second order relation (4), different estimators have the smallest AMSE. In [10] it was shown that the DPR estimator has the same order of the rate of convergence as these four estimators, and it was possible to compare the estimator $\hat{p}$ with all $\gamma_{N,k}^{(i)}$, $i = 1,2,3,4$. The fact that the DPR estimator has the same order of the rate of convergence as these four estimators is not unexpected, since parameter m in estimator $\hat{p}$ plays a similar role as k plays in the definition of estimators $\gamma_{N,k}^{(i)}$, and, taking $2m$ terms in total from a sample, we essentially take the largest values.

Although estimators $\gamma_{N,k}^{(1)}$ and estimator $\gamma_{N,k}^{(4)}$ asymptotically perform better than the DPR estimator $\hat{p}$ (they have smaller AMSE for all values of parameters α, β), relation between other two estimators and $\hat{p}$ is the same as in [5]: for some values of parameters α, β the estimator $\hat{p}$ performs better than $\gamma_{N,k}^{(2)}$ and $\gamma_{N,k}^{(3)}$.

The next steps to improve the DPR estimator were made in [13] and [11]. In [13] the ideas of the Hill and the DPR estimators were successfully combined. At first, the tail index estimation procedure is the same as for the DPR estimator, division of a sample in n groups with m elements in each group. Then, instead of taking two largest elements in each group, Qi takes Hill estimator in each group, using $s+1$ $(1 \le s \le m-1)$ largest values from the ordered statistics $M_{n,i}^{(1)} \ge \dots M_{n,i}^{(m)}$ in each group V_i. Averaging these estimators over groups Qi obtains the following estimator of the parameter $\gamma = \alpha^{-1}$:

$$\gamma_N(s) = \frac{1}{ns} \sum_{i=1}^{n} \sum_{j=1}^{s} \left(\log M_{n,i}^{(j)} - \log M_{n,i}^{(s+1)} \right). \tag{10}$$

In [13] it is proved that, under second-order condition (4), this estimator is asymptotically normal, and that estimator $\gamma_N(1)$ performs better that the DPR estimator for all $\gamma > 0$.

In [11] all class of estimators, generalizing the DPR estimator is introduced. The idea is to take some function $f : [0,1] \to [0,\infty]$ such that $\mathbf{E}\, f(W)$ exists where W is from (3), then this expectation will be some function of α, depending, of course, on function f. Let us denote $h(\alpha, f) = \mathbf{E}\, f(W)$. If $h(\alpha, f)$ is a one-to-one map from $[a,b]$ to $[c,d]$ with $[a,b]$ and $[c,d]$ being subsets of $[0,\infty]$, then estimating the quantity $h(\alpha, f)$ and taking the inverse function we get an estimator for α (with the restriction $a \le \alpha \le b$ if $0 < a < b < \infty$). Therefore it is natural to consider statistic of the form

$$\frac{1}{n} \sum_{i=1}^{n} f(\upsilon_{n,i}). \tag{11}$$

In [11] the following family of functions, continuous in the parameter r,

$$f_r(x) = \frac{1 - x^r}{r}, \ -\alpha < r < \infty, \ r \ne 0 \quad \text{and} \quad f_0(x) = -\ln x, \tag{12}$$

is investigated in detail. Let us note that for this family of functions $h_r(\alpha) := h(\alpha, f_r) = \alpha(r+\alpha)^{-1}$. The estimator $\hat{p}$ is obtained taking $f(x) = x$, the estimator $\gamma_N(1)$ from (10) corresponds to the case $f_0(x) = -\ln(x)$ and $h_0(\alpha) = \alpha^{-1}$. Taking f_r in (11), we have the following family of estimators:

$$\hat{p}_r = \frac{1}{n} \sum_{j=1}^{n} \tilde{\upsilon}_{n,r,j}, \ \tilde{\upsilon}_{n,r,j} = \frac{1 - (\upsilon_{n,j})^r}{r}, \ -\alpha < r < \infty, \ r \ne 0\ , \ \tilde{\upsilon}_{n,0,j} = -\ln \upsilon_{n,j}. \tag{13}$$

Taking the inverse of $h_r(\alpha)$, we have estimators $\hat{\alpha}_r = (1 - r\hat{p}_r)\hat{p}_r^{-1}$, $-\alpha < r < \infty$. Considering the estimator $\hat{p}_r$ we additionally assume that $F(1) = 0$. This assumption was used in the paper [13], too, and is needed only for the reason that we consider negative powers of a random variable taking values in the interval $[0,1]$. The main result in [11] is the following theorem.

Theorem 2 [11]. *Let us suppose that $F(1) = 0$ and let F satisfy (4) with $0 < \alpha < \beta < \infty$ and $C_2 \ne 0$. Also, let us suppose that a sequence $m \equiv m(N) \to \infty$, as $N \to \infty$, is such that*

$$N/m^{1+2\zeta} \to \lambda^2 \in [0,\infty). \tag{14}$$

We recall that $\zeta = (\beta - \alpha)/\alpha$. Then

$$\sqrt{n}\,(\hat{\alpha}_r - \alpha) \xrightarrow{\mathrm{D}}_{N\to\infty} (\alpha + r)^2 \mathcal{N}(\mu_r \lambda, \sigma_r^2), \tag{15}$$

where

$$\mu_r = \frac{C_2 \beta \zeta \Gamma(\beta/\alpha)}{C_1^{\beta/\alpha}(\alpha + r)(\beta + r)}, \quad \sigma_r^2 = \frac{\alpha}{(\alpha + r)^2(\alpha + 2r)}. \tag{16}$$

In the same paper it was shown that there is an optimal (in the sense of AMSE) value r_*, which depends on α and β from (4). Moreover, an explicit expression of this optimal value r_* was obtained, namely,

$$r_* = -\frac{1}{2}\left((\alpha+\beta) - \sqrt{(\alpha+\beta)^2 - 2\alpha^2}\right).$$

Since estimators $\hat{p}$ and $\gamma_N(1)$ are in the same family of estimators with $r = 1$ and $r = 0$, respectively, this means that estimator $\hat{\alpha}_{r_*}$, performs better than estimators $\alpha_N(1) = 1/\gamma_N(1)$ and $\hat{\alpha} = \hat{p}(1-\hat{p})^{-1}$. Comparison with the Hill and the Pickands estimators is also given. Again the estimator $\alpha_{N,k}^{(1)} = (\gamma_{N,k}^{(1)})^{-1}$, obtained from the Hill estimator dominates estimator $\hat{\alpha}_{r_*}$ for all possible values of α, β, but the ratio of AMSE of these two estimators for big values of α is close to 1 (see Fig. 2 and Fig. 3 in [11]). Although it is difficult to believe that among generalized DPR estimators it will be possible to find estimator which would dominate Hill estimator for all possible values of α, β, preliminary calculations show that it is possible to construct DPR type estimators, dominating Hill estimator in some regions of values α, β.

These theoretical results allow to state, that DPR type estimators performs quite well and, taking into account simplicity of construction and some other advantages to be listed bellow, deserve future investigation. At present no theoretical results are known for dependent data. In contrast, for traditional estimators, based on rank statistics, this problem is quite well investigated. Another important problem is the choice of parameters m (size of groups) and r_*. We know optimal values of these parameters, but they depend on unknown parameters α, β. At present we can propose only the following two-step procedure. There are known procedures (see, for example, [3]) of estimating of both parameters in the second-order condition (4), thus, as the first step, we estimate these parameters and then use them to calculate m and r_* and apply estimator $\hat{\alpha}_{r_*}$.

One more interesting question is if it is possible to adapt the idea of the DPR estimator for negative γ, that is , for the estimation of the extreme value index.

At the end of this note we would like to discuss the advantages of the DPR estimator and its modifications; these advantages already had been mentioned in [8],[13],[10]. This estimator is well adapted for high frequency financial data (the so-called stamped transaction-by-transaction or tick-by-tick data), when one has a big flow of data (millions of data in short period), since the statistics is very simple and can be calculated recursively, what allows the so-called on-line estimation. Demonstration of such estimation is described in [5], where some procedures for choosing the size of groups m is also given. As it was noted in [13], there are situations,where only few largest values of observations in the blocks are available, then the DPR type estimators still can be applied, while estimators, based on ordered statistics are not applicable. Also it can be noted that the DPR type estimators are well adapted for detecting the change in tail index, since the construction of such estimators allow to keep time structure of arriving data, which is important issue in the change point problem. Moreover, due to the construction of the generalized DPR estimators, the problem of detecting the change in tail index is reduced to the problem of change of mean value of some random variable. Namely, having a sample $X_i,\ i = 1, 2, \ldots, N$, where the index i can be attributed to time when the data X_i is obtained, we get a new sample $Y_j,\ j = 1, 2, \ldots, n$, where $Y_j = f(v_{n,j})$. Now the problem is to decide if there is a change of mean in this new sample. And this problem is well investigated, a lot of results are available. Contrary, the initial problem of the change in the tail index has attracted attention of not so many statisticians, see [7] and references therein and [14] where financial crisis in some Asian financial markets in 1997 was explained by the tail index change in some financial time series. In both papers tests, based on the Hill estimator were used. In [4] the DPR estimator was used for detecting change in the tail index and the results were applied to analysis of the same financial time series from some Asian countries as in [14].

Very recently, the author with his students specializing in financial mathematics had examined the data of several popular financial indices, such as Dow Jones, NASDAQ, $S\&P500$, and NTSE. The analysis was carried for the absolute values of returns, taking the daily data for the period 2005.09.07-2009.08.26 (this period gives $N = 1000$). The generalized DPR estimators from [11] and the same method in detecting the change of a mean as in [4] were used for analysis. It is interesting to note that considering the above written period, for most indices the test did not show that there was a change in the tail index of the distribution of returns, while taking smaller interval 2007.06.21-2009.08.26 for all indices the test shows the change in the tail index. This can be explained by the fact that the change of a mean is detected better if the change is close to the middle point of the interval under consideration. Since financial crisis was in the Autumn of 2008, the first interval 2005.09.07-2009.08.26 is very unsymmetrical with respect to this data, while the second interval is chosen in such a way that the possible time of a change would be in the middle of the interval. All these results will be published elsewhere soon.

REFERENCES

1. Yu. Davydov and V. Paulauskas and A. Račkauskas, More on P-stable convex sets in Banach spaces, *J. Theoret. Probab.*, (2000), **13**, 39–64.

2. A.L.M. Dekkers and J.H.J. Einmahl and L. de Haan, A moment estimator for the index of an extreme-value distribution, *Ann. Statist.*,(1989), **17**, 1833–1855.

3. M.I. Fraga Alves, M.I. Gomes and L. de Haan , A new class of semiparametric estimators of the second order parameter, *Port. Math.*, (2003) **60**, 194–213.

4. K. Gadeikis and V. Paulauskas , On the estimation of a change point in a tail index, *Lith. Math. J.*, (2005) **45**, 272–283.

5. L. De Haan and L. Peng, Comparison of tail index estimators, *Statist. Neerlandica*,(1998), **52**, 60–70.

6. B.M. Hill, A simple general approach to inference about the tail of a distribution, *Ann. Statist.*, (1975), **3**, 1163–1174.

7. M. Kim and S. Lee, Test for tail index change in stationary time series with Pareto-type marginal distribution, *Bernoulii*, (2009), **15**, 325–356.

8. N. Markovich, , *Nonparametric Analysis of Univariate Heavy-Tailed Data*, Jon Wiley & Sons, Chichester, (2007).

9. V. Paulauskas, A new estimator for tail index, *Acta Appl. Math.*, (2003), **79**, 55–67.

10. V. Paulauskas and M. Vaičiulis, Once more on comparison of tail index estimators, *Preprint*, (2010), arXiv:1104.1242.

11. V. Paulauskas and M. Vaičiulis, Some new modifications of DPR estimator of the tail index, *Lithuanian Math. J.*, (2011), **51**, 36–50.

12. J. Pickands, Statistical inference using extreme order statistics, *Ann. Statist.*, (1975), **3**, 119–131.

13. Y. Qi, On the tail index of a heavy tailed distribution, *Ann. Inst. Statist. Math.*,(2010) **62**, 277–289.

14. C. Quintos, Zh. Fan, and P.C. Phillips, Structural change tests in tail behaviour and the Asian crisis, *Rev. Econom. Stud.,* (2001), **13**, 633–663.

15. C. Zhou, A two-step estimator of the extreme value index, *Extremes*, (2008), **11**, 281–302.