A shift-optimized Hill-type estimator

Éva Rácz and János Kertész
Department of Theoretical Physics, Budapest University of Technology and Economics, Budapest, Hungary

Zoltán Eisler
Nimbus Volatility Arbitrage, Capital Fund Management, Paris, France

(Dated: May 19, 2009)

A wide range of natural and social phenomena result in observables whose distributions can be well approximated by a power-law decay. The well-known Hill estimator of the tail exponent provides results which are in many respects superior to other estimators in case the asymptotics of the distribution is indeed a pure power-law, however, systematic errors occur if the distribution is altered by simply shifting it. We demonstrate some related problems which typically emerge when dealing with empirical data and suggest a procedure designed to extend the applicability of the Hill estimator.

PACS numbers: 02.50.Tt, 05.45.Tp, 89.65.Gh

I. INTRODUCTION

Heavy-tailed distributions emerge in many situations, with examples ranging from social networks to earthquake intensities, city sizes etc. (for further examples, see [1] and references therein). In mathematical terms, a random variable X has a heavy upper tail if the probability

$$\mathbb{P}\left(X \geq x\right) \propto x^{-\alpha} L\left(x\right), \tag{1}$$

with $\alpha > 0$ and $L\left(x\right)$ being a slowly varying function of its argument. The function $\bar{F}\left(x\right) \equiv \mathbb{P}\left(X \geqslant x\right)$ is called the (complementary) cumulative distribution function (cdf in the following). The exponent α, termed the *tail exponent* is a parameter of practical importance, since this is the quantitity which determines the frequency of extreme events (e. g. huge losses on the stock market).

The general problem can be formulated as follows: given some finite sample $\mathcal{S} = \{X_1, X_2, \ldots, X_N\}$ of independent, identically distributed elements, of which the distribution can be described by Eq. (1), we intend to find an efficient procedure to estimate the tail exponent. There exist many such estimators, each of those has its advantages and drawbacks. The difficulties lie generally in the following:

- The small x form of the cdf, which in Eq. (1) is incorporated in $L\left(x\right)$, can shorten the “effective tail length”, i. e., the domain where the distribution is close to a power-law. Therefore, the actual form of $L\left(x\right)$ affects the speed of convergence of any estimator.
- If α is relatively large, one needs a huge dataset to have enough points in the tail.
- Linear transformations of a random variable do not affect its asymptotic behavior (i. e., α), but can affect the value of an estimator over a finite sample.

The popular Hill estimator [2] (HE in the following) is based on the n largest observations in the sample, and is defined as follows:

$$\hat{\alpha}_{\mathrm{H}}\left(\mathcal{S}, n\right) \equiv \left[\frac{1}{n-1}\sum_{j=1}^{n-1} \ln\left(\frac{X_{(j)}}{X_{(n)}}\right)\right]^{-1}, \tag{2}$$

with $X_{(1)} \geq X_{(2)} \geq \cdots \geq X_{(N)}$ being elements of the order statistics. Note that Eq. (2) is invariant to multiplication ($\hat{\alpha}_{\mathrm{H}}\left(a \cdot \mathcal{S}, n\right) = \hat{\alpha}_{\mathrm{H}}\left(\mathcal{S}, n\right)$), yet not shift-invariant ($\hat{\alpha}_{\mathrm{H}}\left(\mathcal{S} + s, n\right) \neq \hat{\alpha}_{\mathrm{H}}\left(\mathcal{S}, n\right)$). For a fixed n, the HE is a maximum likelihood estimator of the tail exponent, but the appropriate choice of the tail length n remains an issue, since $\hat{\alpha}_{\mathrm{H}}$ is typically very sensitive to it. The standard way to determine the threshold $x_0 \equiv X_{(n)}$ is to construct a so-called Hill plot, which is $1/\hat{\alpha}_{\mathrm{H}}$ as a function of n, and look for a plateau in the graph (for other evaluation methods, see [3]).

In a recent publication [4], Clauset *et al.* suggest a procedure (CSNE in the following) to solve the former problem, i. e. to find the optimal n, in an automated fashion. This method provides superior results if $L\left(x\right) = \mathrm{const.}$, but inherits the sensitivity of the Hill estimator to the actual form of $L\left(x\right)$.

The shifted Pareto tail (Eq. (3)) presents a special case of Eq. (1):

$$\bar{F}\left(x\right) \propto \left(x+s\right)^{-\alpha} \equiv x^{-\alpha} \cdot \left(1 + s/x\right)^{-\alpha}. \tag{3}$$

Although, at first sight, the introduction of data shifts does not seem a drastic change, it limits the applicability of both the HE and the CSNE procedures.

A method to tackle the problem of data shifts is for example the Fraga Alves estimator [5], which is invariant to both shifts and multiplication of the dataset, at the cost of a slow convergence. Another example is the Meerschaert–Scheffler estimator [6] which is shift-independent, but not invariant to multiplication, and furthermore its applicability is restricted to $\alpha < 2$.

The aim of the present work is to show an extension of the Hill estimator which can handle both the threshold and the shift problem, in a similar procedure as CSNE. The paper is organized as follows: In Section II we analyze the systematic errors introduced by the shift in the distribution and

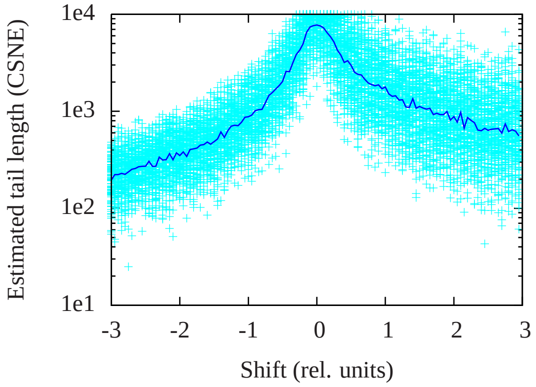

Figure 1: Effective tail length as a function of data shift. Data details: $\alpha = 2.36$ pure power-law, shifts are given in units of the mean absolute deviation ($\mathbb{E}\left(|X - \mathbb{E}\left(X\right)|\right) = \frac{2}{\alpha-1}(\frac{\alpha}{\alpha-1})^{-\alpha+1} \approx 0.7$). For each shift, 100 samples of size 10000 were generated, the cyan colored points correspond to the result of the fitting procedure on each of those. The blue colored curve corresponds to the mean of the 100 runs at each shift. (color online)

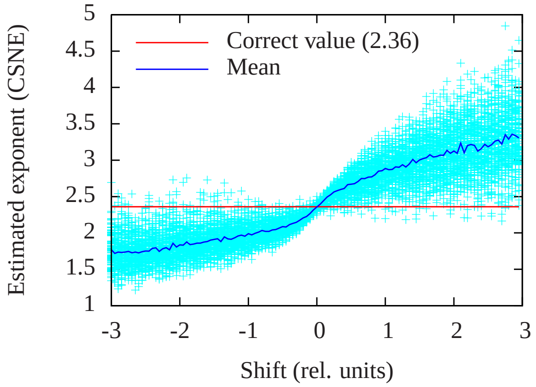

Figure 2: Estimated exponents as a function of data shift, same setup as in Fig. 1. The blue colored curve corresponds again to the average of the estimates. Note that shifts result not only in a shorter tail, thus a larger standard deviation, but also in the shift of the mean estimate. (color online)

demonstrate how this can be taken into account in the estimator. Section III demonstrates the performance of the suggested method on computer-generated data. In Section IV we present an analysis of empirical data as taken from traded volumes of stocks on the stock market. In Section V we give the conclusions, the Appendix briefly summarizes the CSNE algorithm [4].

II. OPTIMIZING THE SHIFT

In short, the CSNE algorithm [4] optimizes the parameter n of Hill's estimator, so that the distance of the fitted power-law and the conditional cdf be minimal (for a summary in terms of formulas, see the Appendix). Clauset *et al.* suggest the Kolmogorov–Smirnov (KS in the following) statistic ($D_{\mathrm{KS}}\left(F, G\right) \equiv \sup_x \left|F\left(x\right) - G\left(x\right)\right|$) as the definition of distance. This method has proven to be a useful tool, however, tests on shifted power-law samples (Eq. (3)) show that in that case it provides biased results (Figs. 1 and 2). While shifts are irrelevant asymptotically (if $|s| \ll x$, then $\left(x+s\right)^{-\alpha} \approx x^{-\alpha}$), having a finite dataset at hand they result in a shorter "effective tail length" (i. e. the threshold above which the $\left(x+s\right)^{-\alpha} \approx x^{-\alpha}$ approximation is valid becomes higher). This observation explains that with growing shift, the CSNE procedure provides more and more volatile results (Fig. 2). The average estimate deviates from the zero shift value since in the shifted case, the cdf is deterministically either convex or concave, i. e. the estimator is bound to deviate from the true value in a fixed direction. Thus, the task is to optimize the tail length n and the shift parameter s simultaneously. If this is achieved, the Hill formula can be applied to the n largest observations shifted with the previously obtained value of s.

In the simplest case, let us assume that whole the sample is taken from a shifted power-law distribution, i. e., we do not have to deal with estimating the tail length. Our aim is to find a shift estimator $\hat{s}\left(\mathcal{S}\right)$, for which $\mathcal{S}' = \mathcal{S} + \hat{s}\left(\mathcal{S}\right)$ is well-approximated by a pure, non-shifted Pareto law. Note that, from the practitioner's point of view, $\hat{s}$ does not necessarily need to be very accurate, since as Fig. 2 shows, the mean estimate depends smoothly on the shift and has a small standard deviation in the vicinity of zero shift.

In geometric terms, we have to "straighten out" the cdf plot, i. e., to determine the shift so that the cdf of $\mathcal{S}'$ on a doubly logarithmic plot is as close to a straight line as possible. The simplest way to achieve this is to minimize the mean squared error of the linear fit on the log-log plot (via numerical optimization, e. g. the golden section method [7]). Figure 3 and Table I show the performance of the latter procedure on computer-generated shifted Pareto samples. It can be concluded that although this type of estimator slightly underestimates the shift on average, it provides reasonable results.

α	$\langle\Delta\hat{s}\rangle$	$\sqrt{\langle\Delta\hat{s}^2\rangle}$
1.5	-0.018	0.070
1.75	-0.019	0.070
2.0	-0.020	0.072
2.25	-0.021	0.074
2.5	-0.023	0.076

Table I: As the table shows, the procedure depends on α, its bias and standard deviation both increase slightly with growing α. (The averages were taken over all shifts for a fixed exponent, i. e. 19x1000 trials with sample size $N = 10^4$.)

Back to the general case, where the tail length can be smaller than the sample size, one has to estimate n along with the shift s. A consistent and simple way to incorporate the shift estimator in a procedure in the manner of [4] is to estimate at each tail length n the shift s based on the n largest entries of the sample. Having obtained s, one has to shift the tail with this value, and calculate the Hill es-

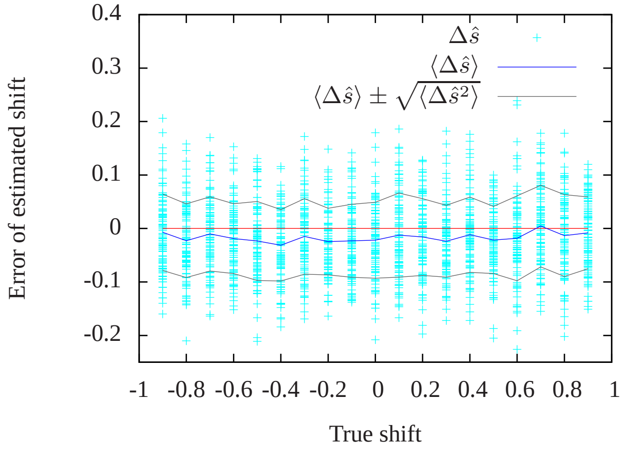

Figure 3: The figure shows the shift estimator's output on computer-generated shifted power-law samples ($\alpha = 2.0$, $N = 10^4$), at each shift value, 100 trials were performed. Note that the procedure is definitely shift-invariant, as intended. (color online)

timator using the shifted tail. Thus, we obtain (n_i, s_i, α_i), $i = 1, 2, \ldots, K \leqslant N$, and from these, we accept the one which is the closest to the empirical cdf, according to the KS statistic. So the "recipe" is the following:

1. For each tail length n, calculate $\hat{s}(\mathcal{T}_n)$ (with $\mathcal{T}_n$ denoting the n largest elements in the sample),
2. calculate $\hat{\alpha}_{\mathrm{H}}(\mathcal{S} + \hat{s}(\mathcal{T}_n), n)$,
3. calculate the KS-distance between the cdf of $\mathcal{T}_n$ and $\left(\frac{x+\hat{s}}{x_0+\hat{s}}\right)^{-\alpha_{\mathrm{H}}}$, with $x_0 \equiv X_{(n)}$, as previously.
4. Accept the fit with the lowest KS-statistic.

The set of n tail lengths to test is chosen in the same manner as in the CSNE procedure (see the Appendix). Sections III and IV analyze the capabilities of this procedure on computer-generated and empirical data. When considering empirical data, one cannot assume that the cdf has exactly the form of Eq. (3), rather a variant of Eq. (1):

$$\bar{F}(x) \propto (x+s)^{-\alpha} \cdot \tilde{L}(x) \equiv x^{-\alpha} \cdot L(x). \tag{4}$$

Although the difference between Eq. (1) and (4) is only in grouping, it can pay off in case $\tilde{L}(x)$ is closer to a constant than $L(x)$.

III. SIMULATION RESULTS

The procedure was tested on computer-generated datasets of size $N = 10^4$ consisting of independent elements distributed according to a shifted power-law:

$$\mathbb{P}(X_i \geqslant x) = \begin{cases} (x+s)^{-\alpha} & \text{if } x \geqslant 1, \\ 1 & \text{otherwise,} \end{cases} \tag{5}$$

α	$\langle\hat{\alpha}\rangle$	$\sqrt{\langle\Delta\hat{\alpha}^2\rangle}$
1.5	1.48	0.08
2.0	1.97	0.11
2.5	2.46	0.16

Table II: Average and standard deviation of the exponents estimated with the new method.

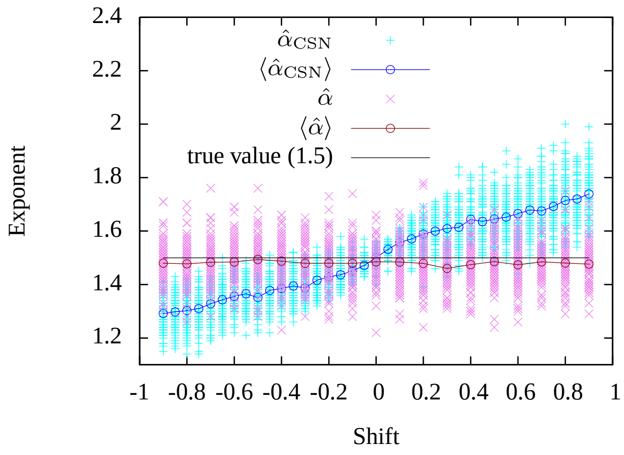

Figure 4: Comparison of the CSNE estimator and the method introduced in Section II. Note that the standard deviation obtained using this new procedure is larger than that of the CSNE at zero shift, this is the price of shift-independence. (color online)

for $i = 1, 2, \ldots, N$. The parameters s and α were varied in the range

$$\begin{aligned} s &\in [-0.9, 0.9], \\ \alpha &\in \{1.5, 2, 2.5\}. \end{aligned}$$

Figures 4–5 show the output of the procedure introduced in Section II for datasets with a fixed exponent $\alpha = 1.5$. One can conclude that the method accounts for the shift problem, although at the price of an increased standard deviation relative to the CSNE zero shift case. The estimates of the shift and the tail length are not as accurate as those of the exponent, but this is not surprising, as the tail-exponent is relatively stable to small changes in the shift and the tail length as well.

Table II shows the performance of the new method for different α values, the averages comprise the estimates with all shift-values considered. The accuracy gets worse with increasing exponent, this is no surprise, since the shift estimator and the Hill estimator both display this property.

IV. EMPIRICAL DATA

As an application, we use the method to determine tail exponents of stock market data. The dataset considered was that of the dollar value [13] of individual transactions of the 1000 largest stocks (according to the total number of trades in the period) in 1994-95 at the New York Stock Exchange [8]. We

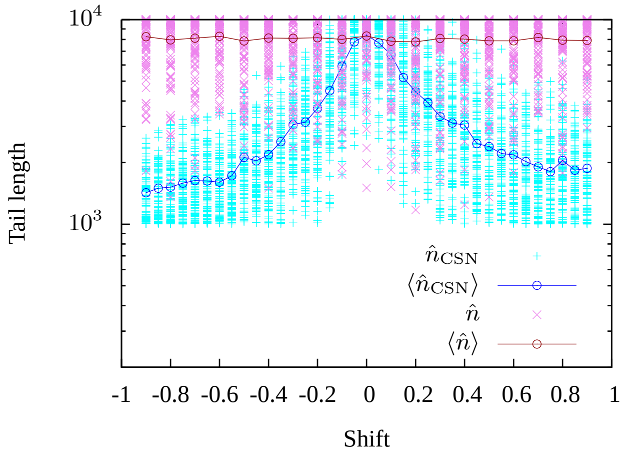

Figure 5: The estimated tail sizes in the procedure introduced in Section II. (color online)

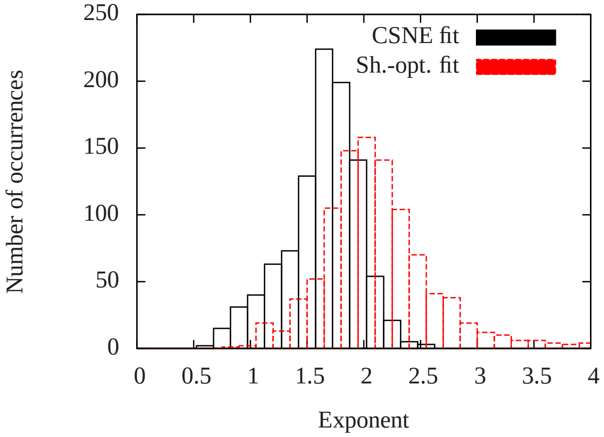

Figure 6: Comparison of the CSNE and the new procedure on empirical data. The black bars correspond to the histogram of the CSNE estimates, the red dashed bars to the new one. (color online)

wish to emphasize, however, as there is no firm theoretical evidence supporting the power-law property of the distribution of trade sizes, that we see it only as one possibility to model the tail of the cdf. Fig. 6 compares the histogram of the tail exponent estimates gained using the CSNE estimator, and the new, shift-invariant procedure. Fig. 7 shows the fits provided by the two estimators on a stock for which they provide very different estimates.

The new procedure found fits with a KS distance on average 23% lower than the CSNE estimator. This is not a drastic difference, but on a logarithmic scale, the KS statistic is less sensitive in the tail than for small x values. Fig. 7 demonstrates this in the case of a specific stock (Kmart Corporation): although on a linear scale (upper part), the two fits do not seem to be very different, the logarithmic plot (lower part) shows that the two deviate in the tail region. In this specific case, the distance of the new fit improved from 0.035 to 0.021.

In empirical data, an additional problem has to be considered when analyzing the results. In case there is no strong theoretical indication for Eq. (4), this type of ansatz can only be accepted, if there are at least some datapoints in the tail, i. e., for which the approximation

$$\frac{x^{-\alpha}}{(x+s)^{-\alpha}} \approx 1+\frac{\alpha\cdot s}{x} \tag{6}$$

is applicable. This is of importance because fitting an exponent to a non-observed tail is questionable, and even if there is theoretical evidence for a shifted power-law tail, errors are amplified. In other words, the quantity

$$\delta = \frac{\hat{\alpha}\cdot\hat{s}}{x_{\max}}, \tag{7}$$

(with $x_{\max} = \max_i \{X_i \in \mathcal{S}\}$) which measures the "distance" of the largest observation from the tail region, has to be small. For the data analyzed, about 10% of the stocks had $\delta \geqslant 0.1$. If we exclude these data from the statistic, the average exponent is 2.0 with a standard deviation of 0.35 [14].

One can conclude that in the case of stock trade volumes, the inclusion of data shifts clearly has an effect on the results. Furthermore, note that the typical value of the estimates is approximately 2, i. e. on the boundary of the Levy regime. This matter has been controversial and our present results support the view that the exponents are higher than thought earlier (Refs. [9],[10],[11] and [12]). Furthermore, since the estimates have a large standard deviation, we find that the term universality is not applicable to trade volume distributions.

V. CONCLUSIONS

In this paper, we have shown for empirical as well as computer-generated data, that data shifts can play an important role in the tail exponent estimation procedure. Tests on computer-generated datasets showed that this problem can be solved by the suggested extension of the Hill estimator. A small, yet systematic underestimation of the exponent is present, nevertheless, on empirical data, this bias loses its importance when compared to other error sources. Our results regarding stock market data lead us to doubts about the idea of universal tail exponents (Ref. [9]-[10]) regarding both universality and typical values.

Appendix: THE CSNE ALGORITHM [4]

The CSNE method optimizes the tail length parameter of the Hill estimator: the algorithm calculates $\hat{\alpha}_{\rm H}(\mathcal{S},n)$ for many different n-s and accepts the fit which is the closest to the empirical cdf, according to the Kolmogorov–Smirnov statistic. In terms of formulas, the procedure for a given set $\mathcal{S} = \{X_1, X_2, \ldots, X_N\}$ observations, can be described as follows:

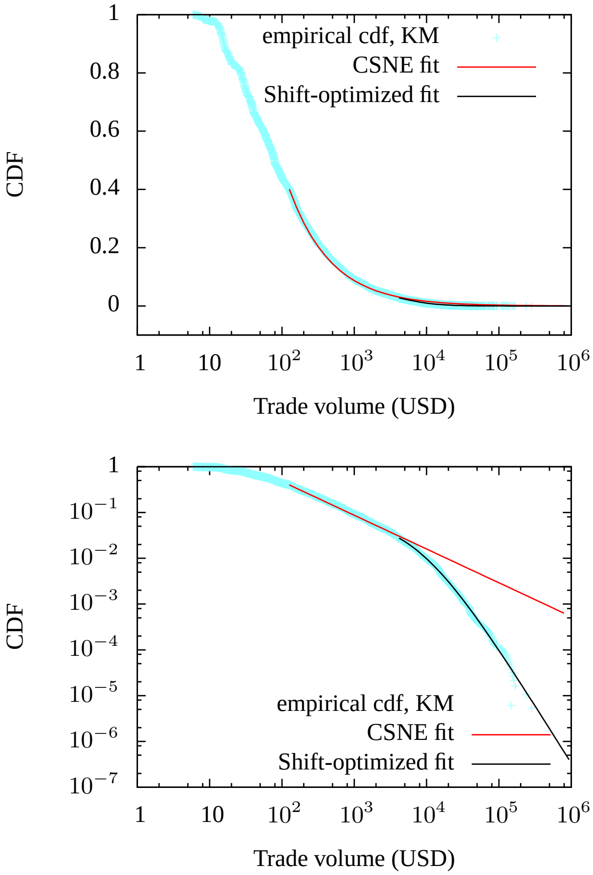

Figure 7: The cdf of trade sizes for the shares of the Kmart Corporation (cyan crosses). The red and black curves show the fits provided by the two estimation procedures. The upper and lower plots differ only in the scale of the vertical axis. (color online)

1. Sorting:

$$\{X_1, \ldots, X_N\} \to \{X_{(1)}, \ldots, X_{(N)}\} : \\ X_{(1)} \geqslant X_{(2)} \geqslant \cdots \geqslant X_{(N)}$$

2. Choose the set $\mathcal{C} \subseteq \{1, 2, \ldots, N\}$ of tail lenghts to check, according to the following criteria:

   - Do not include $n \leqslant n_{\min}$, because it does not make any sense to calculate the Hill estimator based on e. g. 5 elements.
   - If $N - n_{\min} + 1 > M$ choose M elements from $\{n_{\min}, n_{\min} + 1, \ldots, N\}$ uniformly to obtain $\mathcal{C}$.

3. Distance: Kolmogorov–Smirnov statistic:

$$\tilde{F}_x(x_i) \equiv \mathbb{P}(X \geqslant x_i \mid X \geqslant x) \\ D_{\mathrm{KS}}(x) = \max_{x_i \geqslant x} \left| \tilde{F}_x(x_i) - \left(\frac{x_i}{x}\right)^{-\hat{\alpha}_{\mathrm{H}}} \right| \qquad \text{(A.1)}$$

4. For all $X_{(n)} \in \mathcal{C}$ calculate the Hill estimator $\hat{\alpha}_{\mathrm{H}}^{-1}(\mathcal{S}, n)$ and the distance of the Hill fit and the empirical distribution function, $D_{\mathrm{KS}}(n)$.
5. Accept the tail-length n which minimizes the distance:

$$\hat{n} = \arg\min_{n \in \mathcal{C}} D_{\mathrm{KS}}\left(X_{(n)}\right)$$

6. As a result, we obtain:

   - $\hat{\alpha} = \hat{\alpha}_{\mathrm{H}}(\mathcal{S}, \hat{n}_{\mathrm{tail}})$
   - $\hat{x}_0 = X_{(\hat{n})}$
   - $d = D_{\mathrm{KS}}(\hat{x}_0)$

[1] M. E. J. Newman. Power laws, Pareto distributions and Zipf's law. *Contemp. Phys.*, 46:323–351, 2005.

[2] B. M. Hill. A simple general approach to inference about the tail of a distribution. *Ann. Stat.*, 3:1163–1174, 1975.

[3] H. Drees, L. de Haan, and S. Resnick. How to make a Hill plot. *Ann. Stat.*, 28:254–274, 2000.

[4] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *arxiv: 0706.1062v1*, 2007.

[5] M. I. Fraga Alves. A location invariant hill-type estimator. *Extremes*, 4(3):199–217, 2001.

[6] M. M. Meerschaert and H-P. Scheffler. A simple robust estimation method for the thickness of heavy tails. *J. Stat. Plann. Inference*, 71:19–34, 1998.

[7] W. T. Vetterling and B. P. Flannery. *Numerical Recipes in C++: The Art of Scientific Computing*. Cambridge University Press, 2002.

[8] Trades and Quotes Database for 1993-2003, New York Stock Exchange, N ew York.

[9] P. Gopikrishnan, V. Plerou, X. Gabaix, and H. E. Stanley. Statistical properties of share volume traded in financial markets. *Phys. Rev. E*, 62:R4493–R4496, 2000.

[10] X. Gabaix, P. Gopikrishnan, V. Plerou, and H. E. Stanley. Understanding the cubic and half-cubic laws of financial fluctuations. *Physica A*, 324:1–5, 2003.

[11] Z. Eisler and J. Kertész. Size matters: some stylized facts of the stock market revisited. *Eur. Phys. J. B*, 51:145–154, 2006.

[12] V. Plerou and H. E. Stanley. Tests of scaling and universality of the distributions of trade size and share volume: Evidence from three distinct markets. *Phys. Rev. E*, 76:046109, 2007.

[13] We measured trade volumes in US dollars to avoid the problem of stock splits.

[14] Fig. 6 is the histogram of all obtained exponents, $\delta \geqslant 0.1$ included. The average is in the non-filtered case higher, 2.13.