Reply to "Comment on 'Tests of scaling and universality of the distributions of trade size and share volume: Evidence from three distinct markets' "

Vasiliki Plerou and H. Eugene Stanley

Department of Physics and Center for Polymer Studies, Boston University, Boston, Massachusetts 02215, USA

(Received 16 January 2009; published 24 June 2009)

Analyzing trade-by-trade data for three distinct markets, we showed that the cumulative distributions of trade size display power-law tails $P\{q>x\}\sim x^{-\zeta_q}$, with exponents ζ_q in the "Lévy stable domain" ($\zeta_q<\zeta_q^*=2$). Moreover we reported that the exponent values are consistent for all stocks irrespective of stock-specific variables such as market capitalization, industry sector, or the specific market where the stock is traded. Our conclusions were based on using two distinct estimation methods. Rácz *et al.* now propose that one of the estimators we used has slow convergence for a pure power law, particularly as tail exponents approach the boundary $\zeta_q^*=2$. We examine the robustness of our results to specific estimation method by additional analysis using five distinct techniques to estimate ζ_q. We find results that are fully consistent with those we had reported, providing compelling evidence that our conclusions hold regardless of estimation procedure.

DOI: 10.1103/PhysRevE.79.068102 PACS number(s): 89.75.Da, 89.65.−s, 05.45.Tp

I. INTRODUCTION

We recently analyzed tick-by-tick data for three distinct markets: (i) the New York Stock Exchange (NYSE), (ii) the London Stock Exchange (LSE), and (iii) the Paris Bourse, and reported that the cumulative distribution of individual trade size q displays a power-law tail

$$P\{q>x\}\sim x^{-\zeta_q}. \tag{1}$$

Here ζ_q is the tail exponent [1,2]. In addition to evidence from performing power law regressions and obtaining estimates of ζ_q, our analysis provided the following estimates for ζ_q:

(a) U.S. stocks: for the largest 1000 U.S. stocks, during the 2 yr period of 1994–1995, we reported using three different methods [1,2]:

$$\zeta_q=\begin{cases}1.53\pm0.07 & (\text{Hill estimator [3]})\\ 1.45\pm0.03 & (\text{Scaling of moments})\\ 1.63\pm0.03 & (\text{MS estimator [4]}).\end{cases} \tag{2}$$

(b) LSE stocks: for the largest 85 stocks traded on the LSE that are part of the FTSE[1] 100 index during the 2 yr period of 2001–2002, we found [1]:

$$\zeta_q=\begin{cases}1.57\pm0.01 & [\text{Hill estimator}]\\ 1.58\pm0.01 & [\text{MS estimator}].\end{cases} \tag{3}$$

(c) Paris Bourse stocks: for the 13 large stocks that formed part of the CAC 40 index[2] during the 4 yr period of 1995–1999, we found [1]:

$$\zeta_q=1.53\pm0.04\ [\text{Hill estimator}]. \tag{4}$$

Note that all our estimates of ζ_q are within the Lévy stable domain $\zeta_q<\zeta_q^*=2$. Moreover, the values of ζ_q are consistent not only between stocks within the same market but also across the three markets analyzed [1] (cf. the discussion in Refs. [5,6] for a discussion of ζ_q for LSE stocks).

Reference [7] points out that one of the estimators that we use, the Meerschaert-Scheffler (MS) estimator [4], is constructed specifically for the domain $\zeta_q<2$, and is therefore unable to distinguish between the two cases: (i) $\zeta_q<2$ and (ii) $\zeta_q>2$. Our usage of the MS estimator is guided by the range of exponent values $\zeta_q\approx1.45-1.6$ (obtained from alternate methods listed above), which are within the domain of applicability of the MS estimator. In addition, Ref. [7] shows that, for a "pure" power-law distribution, the MS estimator has slow convergence, and underestimates the tail exponents.

In this Comment, we report the following results:

(i) We first apply five additional estimation techniques to measure the exponent ζ_q. Each one of these techniques gives results that are consistent with the values of ζ_q [Eqs. (2)–(4)] as reported in Ref. [1]. Together they give compelling evidence in support of the robustness of our conclusion that $\zeta_q<\zeta_q^*=2$.

(ii) Next we discuss the problem of estimating ζ_Q, the tail exponent describing the *cumulative* distribution function,

$$P\{Q>x\}\sim x^{-\zeta_Q}, \tag{5}$$

of share volume in a time interval Δt

$$Q_{\Delta t}\equiv\sum_{i=1}^{N_{\Delta t}}q_i, \tag{6}$$

where $N_{\Delta t}$ denotes the number of trades in the time interval Δt. We show that, for a control generated using q_i distributed as a power law with ζ_q within the Lévy stable domain, aggregation can give apparent exponent values $\zeta_Q>2$ similar to those observed in Ref. [8].

(iii) Lastly, we discuss the performance of the MS estimator. Although for a pure power-law distribution [e.g., $P(q>x)=a^{\zeta_q}x^{-\zeta_q}$ if $x>a$, and 1 otherwise], the MS estimator converges slowly as illustrated in Fig. 1 of Ref. [7]; we note that its convergence and accuracy is better for genuinely Lévy stable distributions.

[1] See weblink www.ftse.com

[2] See weblinks www.nyse.com/cac40 and http://en.wikipedia.org/siki.CAD_40

II. ESTIMATION OF THE EXPONENT ζ_q FROM FIVE ADDITIONAL ESTIMATION TECHNIQUES

To obtain additional confirmation that ζ_q is robust to the choice of estimation technique, we have further analyzed the data for the 116 most actively traded U.S. stocks using five separate estimation techniques:

A. Technique 1: max-spectrum estimator

The max-spectrum estimator, proposed in Ref. [9], is a unique approach for estimating the tail exponent ζ_q. Consider an *i.i.d.* series X_i with the cumulative distribution $P(X>x) \sim x^{-\alpha}$. The max-spectrum estimator [9] utilizes the relationship that the maximum event observed over a block of size m scales as $m^{1/\alpha}$. We find [10,11]

$$\zeta_q = 1.59 \pm 0.02. \tag{7}$$

B. Technique 2: moment estimator

The moment estimator [12] is a generalization of Hill's original method obtained by including a correction term. We obtain [13]

$$\zeta_q = 1.73 \pm 0.04. \tag{8}$$

C. Technique 3: Fraga-Alves estimator

The Fraga-Alves estimator [14] is a location-invariant estimator based on Hill's original estimator. Alternative techniques such as Pickand's estimator [15] have the same location-invariant property although it is found to display large variance. We find [16]

$$\zeta_q = 1.71 \pm 0.03. \tag{9}$$

D. Technique 4: discrete Hill estimator

The discrete Hill estimator [17] is a variant of Hill's method for discrete data sets. Applying this estimator to the U.S. data, we find [18]

$$\zeta_q = 1.55 \pm 0.02. \tag{10}$$

E. Technique 5: Hill estimator with optimal threshold

The Hill estimator with optimal threshold [17] relies on using the Kolmogorov-Smirnov statistic to determine the optimal estimation threshold. The optimal threshold is chosen such that the probability distributions of the data and the best-fit power law are as similar as possible beyond this threshold. We find [19,20]

$$\zeta_q = 1.69 \pm 0.02. \tag{11}$$

Each of these five distinct estimation techniques yields results that confirm our findings reported in Ref. [1]. Together, they provide compelling evidence that $\zeta_q < 2$.

We conclude with a graphical analysis of the distribution of trade size $P(q>x)$, certainly the most direct method of analyzing the tail behavior (Fig. 1). The solid line corresponds to $\zeta_q = 1.5$ while the dashed line shows the limit $\zeta_q = 2$. We note that (a) the three distributions are mutually compatible, as corroborated by various exponent estimates listed above, and (b) all three display tail exponents which are consistent with $\zeta_q < 2$.

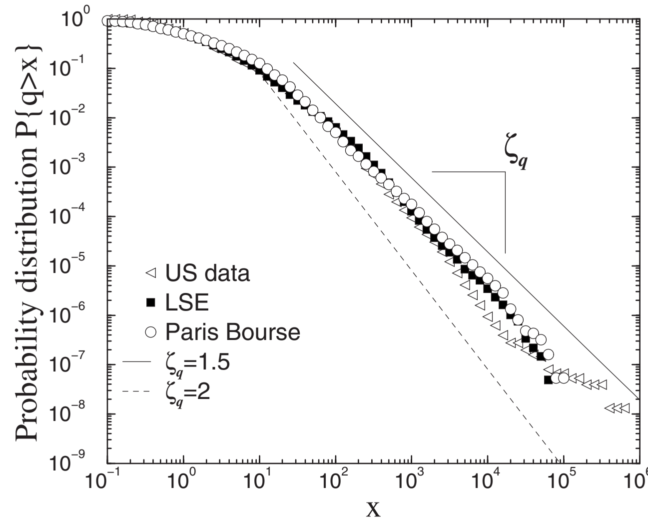

FIG. 1. Distribution function $P(q>x)$ of the trade size q for all three databases analyzed in Ref. [1]. The data for each stock have been scaled by its median. The tails for all three distributions seem consistent with $\zeta_q < 2$ within the Lévy stable domain. Power-law regressions give $\zeta_q = 1.55 \pm 0.04$ (NYSE), 1.73 ± 0.02 (LSE), and 1.60 ± 0.03 (Paris Bourse). Please refer to the text for more accurate exponent estimates.

III. ESTIMATION OF THE EXPONENT ζ_Q

Next we focus on share volume $Q_{\Delta t}$ defined in Eq. (6). Consistent estimation of the exponent ζ_Q, which describes the tail behavior of the cumulative distribution of share volume $P(Q>x)$, is a more challenging problem than estimating ζ_q. First, $Q_{\Delta t}$ acquires long-range correlations from $N_{\Delta t}$ [2], so unlike trade size q, it has pronounced dependencies. Second, estimators such as the Hill estimator which rely on an estimation threshold give biased results (cf. Fig. 6 in Ref. [1]) for aggregated data such as $Q_{\Delta t}$. This estimation bias, which increases with the degree of aggregation (i.e., with increasing Δt), apparently led Ref. [8] to conclude that the "true" exponent ζ_Q increases with increasing Δt with an exponent estimate $\zeta_Q > 2$ implying finite variance. Similar bias is present for other estimators which rely on an estimation threshold, including the "shifted Hill" method, also applied in Ref. [8].

It is instructive to compare the performance of various estimators when applied to computer-generated aggregated data such as

$$Y_{\Delta t} \equiv \sum_{i=1}^{N} x_i, \tag{12}$$

where x_i is generated to follow a power-law distribution with exponent ζ_x [cf. Eq. (16) of Ref. [1]], and N is generated to

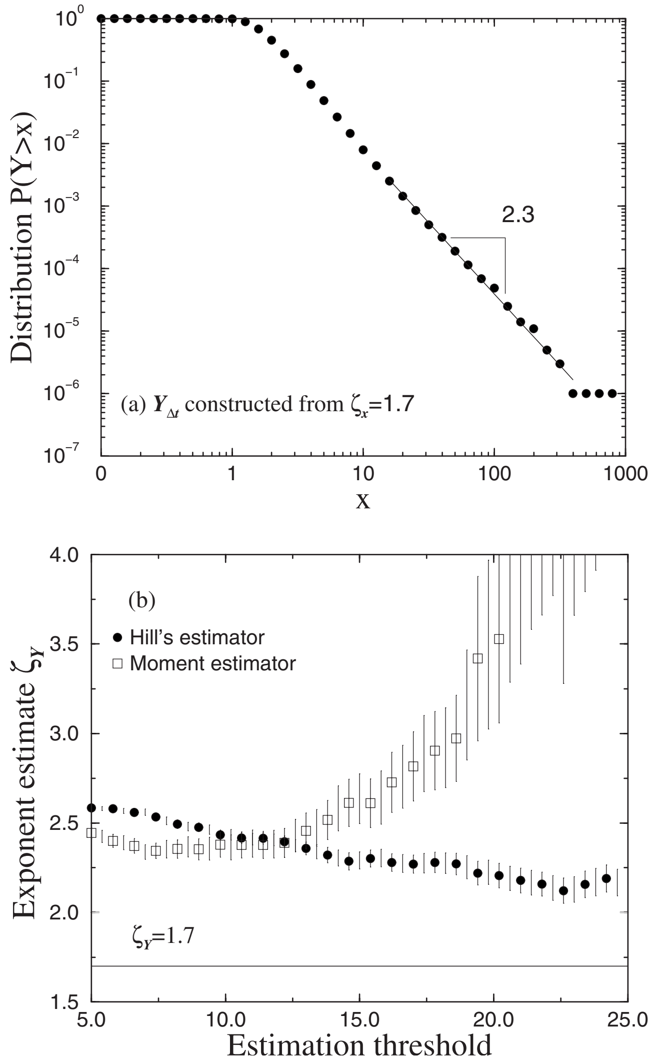

FIG. 2. (a) Distribution function $P(Y>x)$ of aggregated time series $Y_{\Delta t} \equiv \Sigma_{i=1}^{N_{\Delta t}} x_i$, where x_i are generated to have a pure power-law tail of $\zeta_x = 1.7$, $N_{\Delta t}$ is generated to have a pure power-law tail of $\zeta_{N_{\Delta t}} = 3$, and $\langle N_{\Delta t} \rangle$ is set to have the same average value as $\langle N_{\Delta t} \rangle$ for $\Delta t = 15$ min in the NYSE data. The solid line shows a power-law fit which gives the estimate $\hat{\zeta}_Y = 2.28 \pm 0.05$, significantly larger than the true value of $\zeta_Y = 1.7$. (b) Hill estimator and the moment estimator [12] applied to Y as a function of estimation threshold shows an exponent estimate $\zeta_Y > 2$ although the true value is $\zeta_Y = \zeta_x = 1.7$.

have a power law $\zeta_N = 3$ (similar to empirical results [21]). Consider the case when $\zeta_x = 1.7$ (similar to some of the estimates we obtain for ζ_q). Since ζ_x is within the Lévy stable domain, we expect the asymptotic behavior $P(Y>x) \sim x^{-\zeta_Y}$ with $\zeta_Y = \zeta_x = 1.7$ (since $\zeta_x < \zeta_N$) [22]. However, as shown in Fig. 2(a), the distribution function $P(Y>x)$ is virtually indistinguishable from a power law with a larger exponent value—regressions give $\zeta_Y = 2.29 \pm 0.04$.

Figure 2(b) shows the results of applying the Hill's estimator and the moment estimator [12] to the same data. Clearly both of these estimators provide exponent estimates significantly larger than the true value of $\zeta_Y = 1.7$. Note that these estimates, including the power law regression estimate $\zeta_Y = 2.29 \pm 0.04$ are not unlike the estimates of $\zeta_Q \approx 2.3$ reported in Ref. [8]. Indeed, for truly Lévy stable distributions, it is well documented that common estimation techniques such as Hill's estimator give exponent estimates which have significant upward bias when the tail exponent is in the range $1.5 < \zeta_q < 2$ [23]. This fact is particularly relevant in this case since $\zeta_q < 2$, so as N becomes large, the distribution of Q approaches a Lévy stable distribution as claimed in Ref. [1].

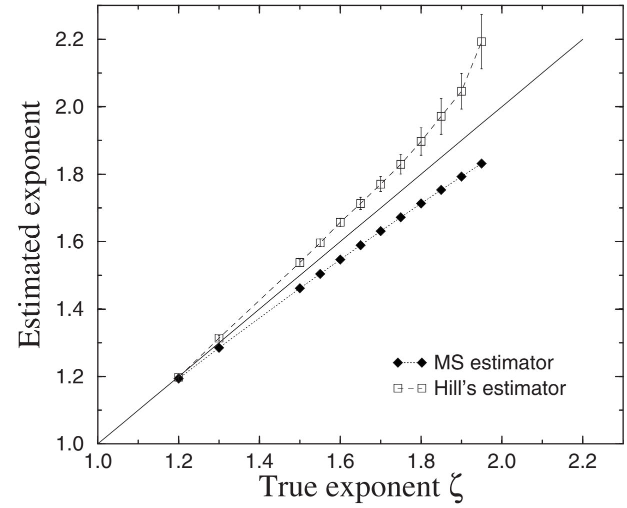

FIG. 3. Comparison of the exponent estimates $\hat{\zeta}$ obtained from the MS and the Hill's estimator for symmetric Lévy stable distributed data generated using the procedure of Ref. [24].

IV. RATE OF CONVERGENCE OF THE MS ESTIMATOR

Reference [7] notes that the MS estimator has slow convergence. We have explored the possibility of using other threshold-independent techniques for estimating ζ_Q, e.g., the Hosking-Wallis (HW) estimator [25]. Although the HW estimator does not have the limitation that its domain of applicability is for exponents <2, it is reliable only for pure Pareto distributions [25] and gives unstable results in the presence of any departures from the pure Pareto form, even at small values. For example, the HW estimator would provide reliable results for control data such as presented in Ref. [7]. In contrast, the MS estimator is more robust to this effect although the rate of convergence is slow. As shown in Fig. 1 of Ref. [7], the deviation between the actual and estimated exponent values is particularly large for the pure power-law form as ζ_q approaches 2 even with 5×10^5 points.

For Lévy stable distributed data however, this deviation is more benign. To illustrate the performance of the MS estimator, we have generated surrogate data following a stable Lévy distribution using the methods of Ref. [24]. Figure 3 shows that the MS estimator underestimates the true exponent while Hill's estimator overestimates. As per the discussion in Sec. III of Ref. [4], the performance of the MS estimator is superior to that illustrated in Ref. [7] because the data in Fig. 3 is generated to follow a Lévy stable law in contrast to a distribution with a power-law tail as in Ref. [7]. We note that, since $\zeta_q < 2$, the distribution of the aggregated data Q would more closely resemble the Lévy stable data in Fig. 3 so the MS estimator, although its convergence is slow, is useful to estimate ζ_Q. In addition, our numerical tests show that the MS estimator is useful in detecting dependencies of the exponent on other variables (more so than Hill's technique), which is particularly relevant in the context of testing the dependence of ζ_Q on variables such as market capitalization.

In summary, we applied five additional estimation techniques to calculate the exponent ζ_q. We first show that each of these techniques give results that are consistent with the values of ζ_q reported in Ref. [1]. Together they give compelling evidence in support of the robustness of our conclusion that $\zeta_q < 2$. Second, we discussed the problem of estimating ζ_Q, the tail exponent describing the decay of the cumulative distribution of share volume $Q_{\Delta t}$. Third, we showed that, for surrogate data generated assuming q_i is distributed as a power law within the Lévy stable domain $\zeta_q < 2$, aggregation can give apparent exponent values $\zeta_Q > 2$.

ACKNOWLEDGMENTS

We thank X. Gabaix for helpful discussions. This reply benefited from the remarks of an anonymous referee.

[1] V. Plerou and H. E. Stanley, Phys. Rev. E **76**, 046109 (2007).
[2] P. Gopikrishnan, V. Plerou, X. Gabaix, and H. E. Stanley, Phys. Rev. E **62**, R4493 (2000).
[3] B. M. Hill, Ann. Stat. **3**, 1163 (1975).
[4] M. Meerschaert and H. Scheffler, J. Stat. Plan. Infer. **71**, 19 (1998); see also M. Herrick *et al.*, Water Resour. Res. **38**, 1227 (2002).
[5] V. Plerou *et al.*, Quant. Finance **4**, C11 (2004).
[6] J. D. Farmer and F. Lillo, Quant. Finance **4**, C7 (2004).
[7] E. Rácz, Z. Eisler, and J. Kertész, previous Comment, Phys. Rev. E **79**, 068101 (2009).
[8] Z. Eisler and J. Kertész, Eur. Phys. J. B **51**, 145 (2006); Phys. Rev. E **73**, 046109 (2006).
[9] S. A. Stoev, G. Michailidis, and M S. Taqqu, e-print arXiv:0609163v1.
[10] The block sizes were chosen in powers of two, and the quoted estimate of ζ_q is obtained by fitting for block sizes m in the range $3 < \log_2(m) < 12$. Restricting the fit to $\log_2(m) > 5$, we obtain $\zeta_q = 1.70 \pm 0.05$. To avoid the effect of serial correlation we have randomized the data before applying this technique.
[11] The error bars listed throughout this paper correspond to one-standard deviation scaled by the square root of the number of observations. Note that this is accurate only for independent data, and tends to underestimate the true magnitude of the error bars. Applying a bootstrap procedure can provide a more reliable estimate of the error bars which indicate values that are $\approx 4-5$ times larger.
[12] A. L. M. Dekkers, J. H. J. Einmahl, and L. De Haan, Ann. Stat. **17**, 1833 (1989).
[13] As with the Hill estimator, this technique also relies on the number of tail events (order statistics). The values provided are using the top 5% of the points. Using the top 10% gives 1.49 ± 0.03 while using the top 2% gives 1.88 ± 0.05.
[14] I. Fraga Alves, Extremes **4**, 199 (2001).
[15] J. Pickands, Ann. Stat. **3**, 119 (1975).
[16] This estimate is obtained by choosing the tail events such that $q > 20$ in normalized units. Choosing a larger threshold $q > 30$ provides an estimate $\zeta_q = 1.63 \pm 0.04$ while choosing a smaller threshold $q > 10$ provides an estimate 1.87 ± 0.04.
[17] A. Clauset, C. Shalizi, and M. Newman, e-print arXiv:0706.1062.
[18] This estimate is obtained using a threshold $q > 10$ (in normalized units of first central moment). Increasing to $q > 20$ gives 1.72 ± 0.03 while $q > 30$ gives 1.81 ± 0.03.
[19] For the 116 stocks in our sample, the average optimal threshold determined by this procedure ≈ 25 (units of first centered moment).
[20] Although the method of Clauset *et al.* tries to find the optimal threshold, as pointed out in Ref. [3], using the Kolmogorov Smirnov statistic may lead to biased results. More accurate estimators which we have not investigated here include H. Drees and E. Kaufmann, Stochastic Proc. Appl. **75**, 149 (1998) and P. Hall, J. Multivariate Anal. **32**, 177 (1990).
[21] V. Plerou, P. Gopikrishnan, L. A. Nunes Amaral, X. Gabaix, and H. E. Stanley, Phys. Rev. E **62**, R3023 (2000).
[22] On this point, note that the convergence properties of a stochastic sum may be different from the deterministic case, cf., S. Mittnik and S. T. Rachev, Econometric Rev. **12**, 261 (1993).
[23] J. H. McCulloch, J. Bus. Econ. Stat. **15**, 74 (1997).
[24] J. Chambers, C. Mallows, and B. Stuck, J. Am. Stat. Assoc. **71**, 340 (1976); See also R. Weron, Stat. Probab. Lett. **28**, 165 (1996).
[25] J. R. M. Hosking and J. R. Wallis, Technometrics **29**, 339 (1987).