Extracting the scaling exponents of a self-affine, non-Gaussian process from a finite-length time series

K. Kiyani,* S. C. Chapman, and B. Hnat

Centre for Fusion, Space and Astrophysics; Department of Physics, University of Warwick, Coventry CV4 7AL, United Kingdom

(Received 25 July 2006; published 28 November 2006)

We address the generic problem of extracting the scaling exponents of a stationary, self-affine process realized by a time series of finite length, where information about the process is not known *a priori*. Estimating the scaling exponents relies upon estimating the moments, or more typically structure functions, of the probability density of the differenced time series. If the probability density is heavy tailed, outliers strongly influence the scaling behavior of the moments. From an operational point of view, we wish to recover the scaling exponents of the underlying process by excluding a minimal population of these outliers. We test these ideas on a synthetically generated symmetric α-stable Lévy process and show that the Lévy exponent is recovered in up to the 6th order moment after only ~0.1–0.5 % of the data are excluded. The scaling properties of the excluded outliers can then be tested to provide additional information about the system.

DOI: 10.1103/PhysRevE.74.051122 PACS number(s): 05.40.−a, 05.45.Tp, 02.70.Rr, 02.50.Ey

I. INTRODUCTION

There is increasing observational evidence that natural systems often show scaling in a statistical sense, coincident with non-Gaussian "heavy tailed" statistics. Complex systems approaches aim to understand these phenomena as universal, with a key quantitative prediction of theory being scaling exponents. Importantly, the identification of universal scaling functions implies the ability to describe many different length and time scales as well as apparently disjoint physical phenomena with the same macroscopic scaling behavior [1–3].

One of the outstanding challenges in complex system science is then to find robust methods that (i) establish whether there is scaling and (ii) accurately determine the scaling exponents for statistical measures of series of data that are of large, but finite length. We seek to determine the scaling properties of probability distributions that are heavy-tailed. The scaling exponents can be determined through the scaling behavior of the moments, usually characterized by computing structure functions. Where the probability density is heavy tailed the moments and structure functions can depend strongly on extremal values, or outliers. Once we insist that the data series is represented by a finite number of measurements, the values at which these outliers occur will always vary between one realization and the next. From an operational point of view, that is, when the underlying behavior is not known *a priori*, these outliers can potentially distort the scaling properties of the data and the values of scaling exponents extracted via the structure functions. In this paper we propose a generic method for excluding these outliers in a manner which does not distort the underlying scaling properties of the data. These outliers also contain information and we explore a method for extracting this. We will test these ideas on numerically generated Lévy processes.

There has been considerable interest in fractional kinetics as providing stochastic models for the data of candidate complex systems [4,5]. Lévy processes and their associated scaling exponents have been identified for example in biological systems (flight time intervals in the foraging of albatrosses [6]), financial markets (price changes in the S&P 500 index [7]), and physical systems (velocity differences in turbulence [9] and atom trapping times in laser cooling [8]). A robust method for determining the Lévy exponent from finite-sized data sets, where the statistics are not known *a priori* is thus important in its own right. The method that we propose here is however quite generic, with application to a wide class of systems that show scaling; for example, those that can be modeled by stochastic differential equations with scaling [10–12]. In this wider context Lévy processes, which have nonconvergent higher order moments, provide a particularly stringent test of our ideas.

Statistical self-similarity. One can characterize fluctuations in a time series $x(t)$ on a given time scale τ in terms of a differenced variable $y(t,\tau)$,

$$y(t,\tau) = x(t+\tau) - x(t), \tag{1}$$

for time t and interval τ, where the time series and/or stochastic process $x(t)$ represents a particular realization or set of observations of the system from which the y's are generated. We consider the case where the $y(t,\tau)$ satisfy the following scaling relation:

$$y(b\tau) \overset{d}{=} f(b) y(\tau), \tag{2}$$

where b is some scale dilation factor; $\overset{d}{=}$ indicates an equality in the statistical and/or distribution sense; f is some scaling function (to be determined); and we have dropped the time argument in the increments y by assuming statistical stationarity. Both b and $f(b)$ are positive. The property in Eq. (2) is a generalized form of *self-affinity,* and in this sense $x(t)$ is a *self-affine* field. Self-affinity is a particular case of statistical self-similarity, i.e., stochastic processes that exhibit the absence of characteristic scales [3,12,13]. We can write the scaling transformations (2) as

*Electronic address: k.kiyani@warwick.ac.uk

 ©2006 The American Physical Society

$$\tau' = b\tau, \quad y' = f(b)y, \tag{3}$$

where the primed variables represent scaled quantities. Conservation of probability under change of variables implies that the probability density function (PDF) of y, $P(y,\tau)$ is related to the PDF of y', $P(y',\tau')$ by

$$P(y,\tau) = P'(y',\tau')\left|\frac{dy'}{dy}\right|, \tag{4}$$

thus giving from Eq. (3)

$$P(y,\tau) = f(b)P'(y',\tau') = f(b)P'(f(b)y,b\tau). \tag{5}$$

The result (5) expresses the fact that the stochastic process $x(t)$ is statistically self-similar, i.e., that a given process on scale τ' (and thus y') maps onto another process based on a different scale τ (and y) by the scaling transformation in Eq. (3); and that the PDFs of both these processes are related by Eq. (5).

We can go further and reduce the expression (5) to a function of one variable. Since the dilation factor b is arbitrary we choose $b=\tau^{-1}$, which gives the important result

$$P(y,\tau) = f(\tau^{-1})P'(f(\tau^{-1})y,1) = f(\tau^{-1})\mathcal{P}_s(f(\tau^{-1})y), \tag{6}$$

and shows that any PDF P of increments y characterized by a time increment τ may be collapsed onto a single unique PDF $\mathcal{P}_s$ of rescaled increments $f(\tau^{-1})y$ and time increment $\tau=1$, by the above scaling transformation. Identification of this unique *scaling function* and the ensuing collapse is a clearer method of discriminating between different (universality) scaling classes than simply identifying the scaling exponents by themselves [1].

In this paper we will consider the scaling as defined by the structure functions. The generalized structure functions of order p are simply defined as

$$S^p(\tau;\pm\infty) = \langle|y|^p\rangle = \int_{-\infty}^{\infty}|y|^pP(y,\tau)dy. \tag{7}$$

The analysis which follows is also valid for the moments; however, structure functions are typically calculated for data. This avoids the result that odd order moments of symmetric PDFs are zero so that as a consequence, in a physical system, they would be dominated by experimental error. Using the transformation (6), the scaling of the structure functions is

$$\begin{aligned} S^p(\tau;\pm\infty) &= \int_{-\infty}^{\infty}|y|^pP(y,\tau)dy \\ &= \int_{-\infty}^{\infty}|y|^pf(\tau^{-1})\mathcal{P}_s(f(\tau^{-1})y)dy \\ &\overset{y'=yf(\tau^{-1})}{=} (f(\tau^{-1}))^{-p}\int_{-\infty}^{\infty}|y'|^p\mathcal{P}_s(y')dy' \\ &= (f(\tau^{-1}))^{-p}\mathcal{S}_s^p(1;\pm\infty). \end{aligned} \tag{8}$$

This formalism encompasses a general class of self-affine systems in the sense that it is not restricted to the well-studied case of monoexponent scaling.

The above result (8) holds provided that the PDF P is defined for all y. However, for finite data sets this is not the case. In this situation we have the integral (7) defined for the interval $[y_-,y_+]$ where the $y_\pm$ are defined in some sense by the largest events measured in the data set. The values of $y_\pm$ will depend on the time scale τ and the sample size N (which will be held constant). Thus the structure functions for the finite data set are

$$S^p(\tau;y_\pm(\tau)) = \int_{y_-(\tau)}^{y_+(\tau)}|y|^pP(y,\tau)dy. \tag{9}$$

Manipulating this in a similar way to Eq. (8) results in the following scaling relation:

$$S^p(\tau;y_\pm(\tau)) = (f(\tau^{-1}))^{-p}\mathcal{S}_s^p(1;y_\pm(\tau)f(\tau^{-1})). \tag{10}$$

If we assume that the largest events $y_\pm$ scale with τ in the same way as the increments y in Eq. (3) [see Eq. (25) later], then Eq. (10) becomes

$$S^p(\tau;y_\pm(\tau)) = (f(\tau^{-1}))^{-p}\mathcal{S}_s^p(1;y_{s\pm}(1)). \tag{11}$$

We will consider the case of self-affine scaling where the scaling function f takes the form of a monoscaling power law $f(b)=b^H=\tau^{-H}$, where H is known as the *Hurst* exponent. Equation (6) then becomes

$$P(y,\tau) = \tau^{-H}\mathcal{P}_s(\tau^{-H}y), \tag{12}$$

and Eq. (8) becomes

$$S^p(\tau;\pm\infty) = \tau^{\zeta(p)}\mathcal{S}_s^p(1;\pm\infty), \tag{13}$$

where $\zeta(p)=Hp$ for this self-affine case. A log-log plot of S^p vs τ for various orders p reveals scaling if present, and the slope of such a plot determines the exponents $\zeta(p)$ [2,14]. One then verifies that $\zeta(p)=Hp$ by plotting $\zeta(p)$ as a function of p.

The aim of this paper is to obtain a good estimate of the scaling properties of Eq. (7), the structure functions at $N\to\infty$, via Eq. (11) for N large but finite; where N represents the length of the time series being considered. However, we can anticipate that simply setting the limits $y_\pm$ of the integral (9) to the largest values found in a given realization of the data, will give a scaling behavior of Eq. (11) which can differ substantially from that of Eq. (13). This problem arises since the y values of the extremal points fluctuate between one realization and the next, and these fluctuations are more significant in heavy tailed distributions. This in turn will strongly modify the integral. We will therefore explore the possibility of choosing a range for the integral (9) based on the scaling property of the data itself, by systematically excluding the most extreme outlying points. This has the added advantage of not requiring *a priori* information about the system.

We stress that as our aim is to extract scaling exponents, we do not attempt to estimate the value of the moments or structure functions. Thus we will not compute an estimate of the integral (7) *per se*, rather we will examine methods for quantifying its dependence on the dilation factor b (or equivalently τ). Hence, our method can be applied to Lévy processes—where the moments are not defined, but where the PDF has scaling.

The paper is organized as follows. We first introduce the Lévy process that we will use to obtain Eq. (9) and briefly survey results pertaining to its asymptotic behavior. We then discuss the effects of finite-sized data sets and demonstrate the effect of removing outliers on the scaling behavior of the Lévy process. We then explore the behavior of these outliers.

II. LÉVY PROCESSES AND FINITE-SIZE EFFECTS

A. α-stable processes

Many stochastic processes exhibit self-affine scaling and are characterized by "broad tails" described by power laws in their PDFs. Some possible mechanisms by which these power laws occur are discussed in [2]. This general class of stochastic processes can be described in the context of so-called *α-stable Lévy processes* [4,15,16]. We will restrict our attention to symmetric α-stable processes. The PDFs L_α^γ of the increments y of these processes are defined through the Fourier transform of their characteristic function

$$L_\alpha^\gamma(y,\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} dk e^{iky} e^{-\gamma\tau|k|^\alpha}, \qquad (14)$$

where $\gamma \geq 0$ and $\tau \geq 0$ are the characteristic scales of the process and describe the width of the distribution; and $\alpha \in (0,2]$ parameterizes the stability of the distribution; α can be heuristically seen as an indication of the variability of the increments of such processes (also known as *Lévy flights*). In this paper we will take $\gamma = 1$ and will consequently reduce the notation L_α^γ to L_α. The form and convention of the parameters in Eq. (14) are similar to that presented in [17]; for a more rigorous discussion of the mathematical properties of such processes readers are referred to [15,16].

From Eq. (14) it follows that the scaling properties of L_α are

$$L_\alpha(y,\tau) = \tau^{-1/\alpha} L_\alpha(\tau^{-1/\alpha} y, 1) = \tau^{-1/\alpha} \mathcal{L}_{s,\alpha}(\tau^{-1/\alpha} y), \qquad (15)$$

from which the Hurst exponent of symmetric α-stable processes is $H = 1/\alpha$, by comparison with Eq. (12). Figure 1(a) shows the $L_\alpha(y,\tau)$ for $\alpha = 1.4$ and a range of $\tau = 2^0, 2^1, \ldots, 2^{10}$; the scaling collapse (15) has been applied to these in Fig. 1(b).

We now focus on the asymptotic behavior of such distributions. By expanding the complex exponential in Eq. (14) and integrating one can show that in the large y limit we obtain the asymptotic behavior

$$\lim_{y\to\infty} L_\alpha(y,\tau) \simeq \frac{\tau\Gamma(1+\alpha)\sin(\pi\alpha/2)}{\pi|y|^{1+\alpha}} = D_\alpha \frac{\tau}{|y|^{1+\alpha}}, \qquad (16)$$

for $y \gg \tau^{1/\alpha}$ [17,24]. It immediately follows that these power-law tails ensure that for the pth moment to exist, $p - \alpha < 0$. Hence the process has no variance defined for $0 < \alpha < 2$, and in the cases where $0 < \alpha \leq 1$ the process will also have no mean defined, i.e., both these quantities and the other higher-order moments are infinite.

A generalized version of the central limit theorem (CLT) [2] ensures that the sum of all independent and identically distributed (IID) random variables with no finite variance that have distributions with power-law tails that go asymptotically as $y^{-1-\alpha}$ $\{\alpha \in (0,2]\}$, will converge to a Lévy distribution of the same index α. In practice, however, we will always obtain a finite mean and variance from a finite-length time series.

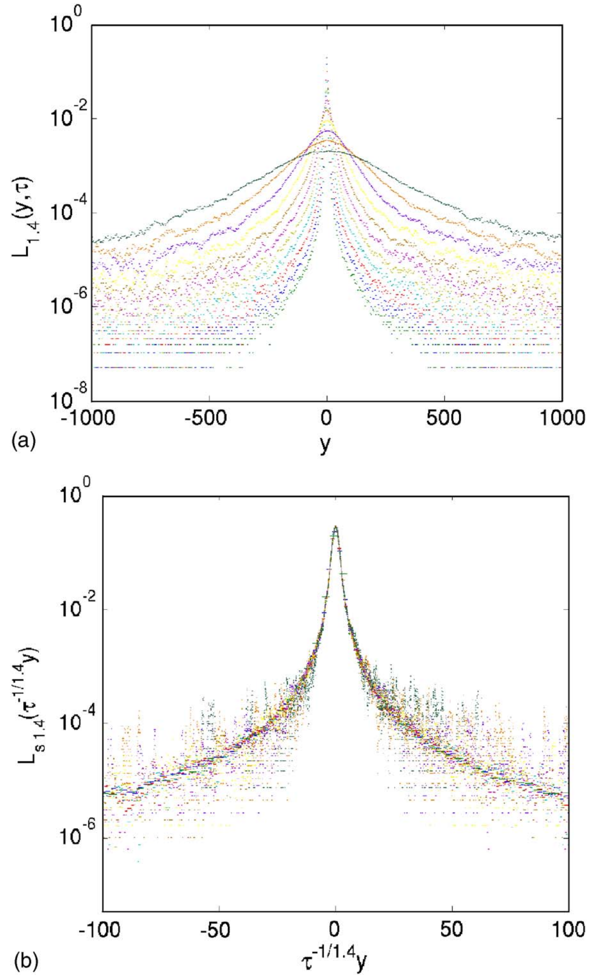

FIG. 1. (Color online) Plots showing probability density functions of the Lévy distribution for index $\alpha = 1.4$ ($N = 10^6$) at different values of differenced interval τ (a) before and (b) after the scaling collapse described by Eq. (15).

B. Finite-size effects and outliers

We will now consider in detail the procedure for extracting the scaling exponents, $\zeta(p)$, from the structure functions in Eq. (13). This centers on first computing $S^p(\tau; y_\pm)$ and the gradients $\zeta(p)$ of log-log plots of $S^p(\tau; y_\pm)$ vs τ. If the process is self-affine [$\zeta(p) = Hp$] we should obtain a straight line on a plot of $\zeta(p)$ vs p from which we can measure the gradient and obtain the Hurst exponent, H. Note that the $\zeta(p)$ for several p are needed to determine H uniquely [12].

However, finite sample sizes result in pseudomultiaffine behavior. As we will show, the primary reason for this anomalous behavior is due to the large scatter in the outlying events of the tails of the distribution. In the case of Lévy-like

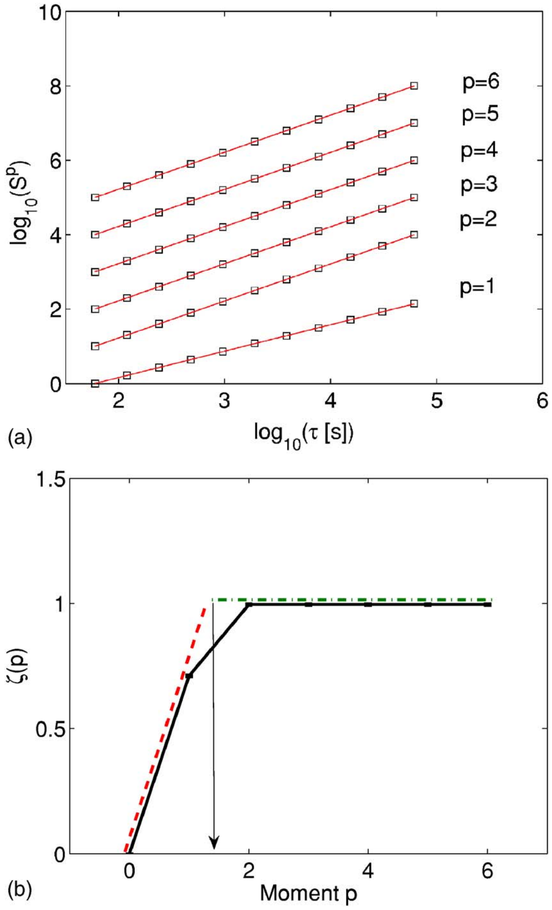

FIG. 2. (Color online) Plots of (a) generalized structure functions S^p vs τ for moments of order $p=1$–6, and (b) the scaling exponents $\zeta(p)$ vs p (solid black line). These quantities are shown for a Lévy process of index $\alpha=1.4$ and with $N=10^6$ data points. The dashed line indicates the expected scaling $\zeta(p)=p/\alpha$ for $p<\alpha$; the dotted-dashed line indicates the scaling exponent observed for $p>\alpha$ in a finite-sized sample. The vertical arrow at $p\simeq\alpha$ separates these two regions of scaling.

processes this scaling bias shows up as a saturation or roll over on the $\zeta(p)$ plots at $p>\alpha$. This can be seen in Fig. 2 which illustrates both the methodology of extracting scaling exponents from structure function plots, and this finite sample size saturation effect in a Lévy process of index $\alpha=1.4$. This saturation effect is well known and an explanation for it can be found in the work by Schmitt *et al.* [5] and Chechkin and Gonchar [18]. We will now establish the scaling properties of these extremal events. We need to emphasize, however, that in contrast to [5,18] we will propose a method for estimating the integral in Eq. (7) such that the scaling in Eq. (13) is recovered for *all* p.

We consider the situation where we have many realizations, that is many data series of size N obtained from the same process. Each of these realizations will have extremal points y^* of their respective PDF. We know the properties of $\bar{y}^*$, the ensemble average of the y^* over the realizations, since it will fall on the Lévy asymptotic distribution (16). We will use a simple example of extreme value theory, EVT (see [2]), to obtain an estimate of the largest event in a sample of N IID measurements of a random variable $y\in\mathbb{R}^+$. An approximation to the probability to see an event that occurs only once can be made by realizing that an event with probability P occurs typically NP times. Therefore, the rarest event in a sample of N measurements, which occurs typically only once, can be seen to be described by $NP(y\geq\bar{y}^*)=1$, where $P(y\geq\bar{y}^*)$ is the probability of observing an event greater than or equal to $\bar{y}^*$; thus

$$P(y\geq\bar{y}^*)=\frac{1}{N}. \tag{17}$$

We can generalize this to the mth largest event:

$$P(y\geq\bar{y}^*_m)=\frac{m}{N}. \tag{18}$$

For the case of the Lévy-like process, within the limits of the integral in $P(y\geq\bar{y}^*_m)$ the main contribution is from the tail and thus we can use Eq. (16) and estimate $P(y\geq\bar{y}^*_m)$ to be

$$P(y\geq\bar{y}^*_m)=\int_{\bar{y}^*_m}^{\infty}L_\alpha(y,\tau)dy\simeq D_\alpha\tau\int_{\bar{y}^*_m}^{\infty}\frac{dy}{|y|^{1+\alpha}}. \tag{19}$$

Evaluating the integral and equating with Eq. (18) gives the following result for the scaling behavior of the mth largest event:

$$\bar{y}^*_m=\left(\frac{D_\alpha N\tau}{m\alpha}\right)^{1/\alpha}. \tag{20}$$

A more detailed account would be to attempt to specify approximately the full PDF of the mth largest event among N IID measurements. Following Sornette [2] the cumulative distribution function (CDF) $\Pi(y<\bar{y}^*_m)$ of the maximum value is

$$\Pi(y<\bar{y}^*_m)=\int_{-\infty}^{\bar{y}^*_m}p_N(y)dy\simeq e^{-(N/m)P(y\geq\bar{y}^*_m)}, \tag{21}$$

where $p_N(y)$ is the PDF of the maximum value among N observations, and is obtained by differentiating Eq. (21) to obtain

$$\frac{d\Pi(y<\bar{y}^*_m)}{dy^*_m}=p_N(\bar{y}^*_m)=\frac{N}{m}L_\alpha(\bar{y}^*_m,\tau)e^{-(N/m)P(y\geq\bar{y}^*_m)}. \tag{22}$$

By substituting Eq. (19) in Eq. (21), we obtain an estimate of the mth largest value, $\bar{y}^*_{m,\Pi}$, that will not be exceeded with probability Π. By setting the left-hand side (LHS) of Eq. (21) to some probability $0<\Pi<1$, we obtain

$$\bar{y}^*_{m,\Pi}=\left(\frac{D_\alpha N\tau}{m\alpha\ln(1/\Pi)}\right)^{1/\alpha}. \tag{23}$$

If one was to set $\Pi=1/2$ the value of y^*_m would correspond to the median value of the mth largest event. To obtain the modal value of $\bar{y}^*_m$, we optimize for the maximum by differentiating Eq. (22) and setting it to zero. This gives us the following solution for the modal value of $\bar{y}^*_m$:

$$\bar{y}^*_{m,mode}=\left(\frac{D_\alpha N\tau}{m(1+\alpha)}\right)^{1/\alpha}. \tag{24}$$

By comparing these expressions one can see that although the approximation of $\bar{y}_m^*$ becomes more refined, the scaling with τ is still that of Eq. (20). Thus we will proceed using the simplest expression (20). In addition, we will be working with a varying fraction m/N rather than varying m or N separately. Importantly, since we are concerned primarily with the scaling with respect to τ, we will write $\bar{y}_m^*$ more informatively as $\bar{y}_m^*(\tau)$ and thus adding to our scaling relations

$$\bar{y}_m^*(\tau) = \tau^{1/\alpha}\bar{y}_m^*(1), \tag{25}$$

as expected from Eq. (2) [26]. We emphasize that this is the scaling of $\bar{y}_m^*$; the average over the mth largest events of a large number of realizations (time series). In practice we will have a single realization and thus one value of y_m^* which will fluctuate about this ensemble averaged $\bar{y}_m^*$. The behavior (25) refers to the property that any point in the curve $P(y,\tau)$ scales as Eqs. (6) and (3).

III. STRUCTURE FUNCTIONS

A. Effects of finite sample size

We can now investigate the scaling behavior of the structure functions of a Lévy-like process, but now with a finite sample size. Following the procedure in Eq. (11) we can discuss the structure functions in the average sense, that is averaged over many realizations of our N sample finite-length time series:

$$\bar{S}^p(\tau;\bar{y}_{1,\pm}^*(\tau)) = \int_{-\bar{y}_{1,-}^*(\tau)}^{\bar{y}_{1,+}^*(\tau)} |y|^p L_\alpha(y,\tau)dy = \int_{-\bar{y}_{1,-}^*(\tau)}^{\bar{y}_{1,+}^*(\tau)} |y|^p \tau^{-1/\alpha}\mathcal{L}_{s,\alpha}(\tau^{-1/\alpha}y)dy, \tag{26}$$

where we have set $m=1$ in $\bar{y}_m^*$ to emphasize that this is the structure function for the raw data with the largest events obviously bounding the data; the subscripts + and − indicate the largest positive and negative events. The substitution $y' = \tau^{-1/\alpha}y$ gives

$$\bar{S}^p(\tau;\bar{y}_{1,\pm}^*(\tau)) = \tau^{p/\alpha}\int_{-\bar{y}_{1,-}^*(\tau)\tau^{-1/\alpha}}^{\bar{y}_{1,+}^*(\tau)\tau^{-1/\alpha}} |y'|^p\mathcal{L}_{s,\alpha}(y')dy' = \tau^{p/\alpha}\left[\int_0^{\bar{y}_{1,+}^*(\tau)\tau^{-1/\alpha}} y'^p\mathcal{L}_{s,\alpha}(y')dy' + \int_0^{\bar{y}_{1,-}^*(\tau)\tau^{-1/\alpha}} y'^p\mathcal{L}_{s,\alpha}(y')dy'\right]. \tag{27}$$

To approximate the integrals in Eq. (27) we assume that values of the largest events are deep in the tail region of the distribution so that we may use the asymptotic form (16). This gives

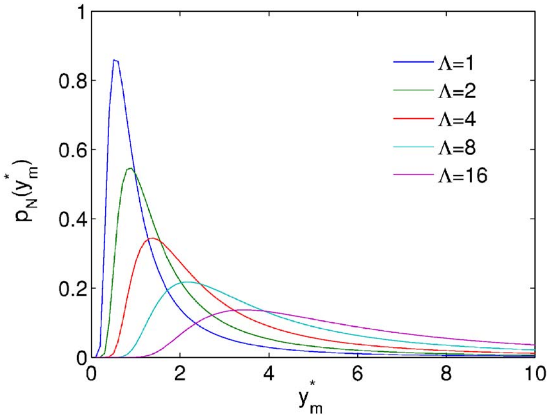

FIG. 3. (Color online) Plot showing the PDF, in Eq. (30), of the mth largest value of a sample size N of a set of measurements taken from a Lévy-like process; the Lévy index $\alpha=1.5$.

$$\bar{S}^p(\tau;\bar{y}_{1,\pm}^*(\tau)) = \tau\frac{D_\alpha}{p-\alpha}\left(\bar{y}_{1,+}^{*(p-\alpha)}(\tau) + \bar{y}_{1,-}^{*(p-\alpha)}(\tau)\right) \quad \forall\, p>\alpha, \tag{28}$$

where the condition $p>\alpha$ is necessary as all structure functions of order $p<\alpha$ of a Lévy distribution exist (i.e., are finite) and this approximation would result in an infrared divergence in Eq. (27), which is clearly incompatible. For the ensemble average Eqs. (19), (20), and (25) hold; thus we can simply substitute Eq. (25) into Eq. (28) to obtain

$$\bar{S}^p(\tau;\bar{y}_{1,\pm}^*(\tau)) = \tau^{p/\alpha}\frac{D_\alpha}{p-\alpha}[\bar{y}_{1,+}^{*(p-\alpha)}(1) + \bar{y}_{1,-}^{*(p-\alpha)}(1)]. \tag{29}$$

In practice the value of y_m^* will vary for each realization of $P(y,\tau)$ about the average $\bar{y}_m^*$ which obeys Eq. (25). For a given functional form of $P(y,\tau)$ the y_m^* will have some probability density $p_N(y_m^*)$ with a statistical spread about the average $\bar{y}_m^*$. An approximation to this can be made by substituting the asymptotic tail form of Eq. (16) into Eq. (22) to obtain

$$p_N(y_m^*) = \frac{\Lambda}{y_m^{*1+\alpha}}\exp\left(-\frac{\Lambda}{\alpha y_m^{*\alpha}}\right), \tag{30}$$

where Λ is given by

$$\Lambda = \frac{ND_\alpha\tau}{m}. \tag{31}$$

Equation (30) is of the form of a stretched exponential. As with any power-law tailed PDF it has infinite variance for $0<\alpha<2$. In the context of EVT, Eq. (30) is not surprising as it is simply an extreme value distribution of type II, i.e., the PDF from a Fréchet distribution. The extreme value distributions can be seen as the large event statistics equivalent to stable distributions (i.e., Gaussian and Lévy). The interested reader is referred to [19,20] for a further discussion of EVT and extreme value distributions.

A plot of the PDF (30) is given in Fig. 3 for various values of Λ and for $\alpha=1.5$. From Fig. 3 we see that as the value of Λ increases, the PDF of y_m^* broadens. Importantly, the PDF of y_m^* (30) has an infinite variance and thus has more frequently occurring extreme values of y_m^* away from $\bar{y}_m^*$. Thus from Fig. 3 and Eq. (31) we see that the scatter in the y_m^* about the average $\bar{y}_m^*$ increases with N and decreases with m/N.

B. Conditioning—overview

We now present a method to "condition" data so that the scaling behavior (13) emerges from the structure functions obtained for a finite data series. From an operational point of view, that is, when attempting to determine an (unknown) exponent from a finite length time series, our aim is to recover Eq. (13) for as many orders p as feasible. This method involves excluding a fraction m/N of the largest events from the data set such that our postexclusion tails are now sufficiently resolved and populated. Although there is some literature on the removal of extreme outliers in data, the first time it was clearly done in the scaling context was by Katul *et al.* [21] followed later by Veltri *et al.* [22,23]. They calculated structure functions via the use of a Haar wavelet transform and conditioned their data by separating the wavelet coefficients into two classes: the majority of coefficients which characterize the "quietly turbulent flow"; and the coefficients which characterize the rare intermittent events corresponding to coherent structures. The partition between these two classes was a wavelet coefficient based upon a multiple F of the square root of the second moment of the coefficients. The easiest way to view this is by looking at the more recent works of Hnat *et al.* [11,12] (and Refs. therein) who employed an equivalent technique but did not use wavelet transforms to calculate the structure functions. Along with their solar wind turbulence data, the latter authors also studied some toy cases of fractional Brownian motion and a Lévy process of $\alpha=1.8$. This conditioning can be succinctly written as the approximation

$$S^p(\tau;\pm\infty)=\int_{-\infty}^{\infty}|y|^pP(y,\tau)dy \to S^C(\tau;\pm A) = \int_{-A}^{A}|y|^pP(y,\tau)dy, \quad (32)$$

where $A=Q\sigma(\tau)$, $\sigma(\tau)$ is the standard deviation, and Q is some constant. This corresponds to *clipping* the wings of the distribution to exclude the very large unresolved events. All of the above studies [12,21,22] showed that removing a relatively few percentage of points is sufficient to regain the scaling. However, the disadvantage of these schemes is that the measure used to exclude the extreme events is the standard deviation, σ, of the *raw* data which must be calculated *a priori* and we have already seen in the above analysis that $p>\alpha$ (and thus σ) is poorly represented in the unconditioned data. A better estimate is to condition the data based on the actual extreme events, i.e., by excluding a certain negligible fraction of the data outliers.

A brief mention should be made of the work by Jespersen *et al.* [24]. They studied the behavior of Lévy flights in external force fields and used a form of conditioning for obtaining a good statistical ensemble in the power-law tail range of a Lévy process. Their conditioning, however, assumes *a priori* knowledge of the distribution and its scaling behavior, and is thus not congruent to the applications to which this paper aims; this being single finite size natural time series.

To summarize, our procedure will be to

(1) Choose limits of the integral in Eq. (32) such that the scaling (13) is recovered—using a method that does not require *a priori* knowledge of the PDF $P(y,\tau)$ to specify those limits.

(2) This procedure will exclude the most outlying points ($\lesssim 1\%$).

(3) These outliers contain some physics of the system. They may or may not share the scaling (12) with the core of the PDF $P(y,\tau)$, instead showing finite-size scaling (exponential rolloff) or other dynamics. Therefore we will also test the outliers for the property (25).

C. Conditioning—Lévy process

We now test these ideas with a numerically generated Lévy process. The increments y of the Lévy process of index α were generated by using the following algorithm [25]:

$$y=\frac{\sin(\alpha r)}{(\cos r)^{1/\alpha}}\left(\frac{\cos[(1-\alpha)r]}{v}\right)^{(1-\alpha)/\alpha}, \quad (33)$$

where $r\in[-\pi/2,\pi/2]$ is a uniformly distributed random variable and v is an exponentially distributed random variable with unit mean. Expression (33) corresponds to the Lévy distribution (14) with $\gamma=1$ and $\tau=1$. We generate a sample of size N and then construct a time series by use of a cumulative sum. This time series was then differenced at various τ as in Eq. (1) using an overlapping window; appropriate here since the data increments are uncorrelated. Structure functions of the increments, $S^p(\tau;y_\pm^*(\tau))$, are then calculated at different orders p and at different values of τ. These are then plotted on a S^p vs τ plot and a linear regression is performed to obtain the gradients $\zeta(p)$ for each moment order p. The plots of these $\zeta(p)$ vs p are shown in Fig. 4 for the two cases $\alpha=1.0$ and $\alpha=1.8$. The error bars in Fig. 4 were obtained from the difference between the linear regression of the structure functions for all values of τ concerned, and the linear regression with the last two τ values not included.

In Fig. 4 we see that if no outliers are removed from the integral for S^p, the resulting values of $\zeta(p)$ for $p>\alpha$ saturate to unity. Removing a small fraction ($\sim 0.001\%$) of the outliers results in a drastic change in the $\zeta(p)$, again emphasizing the strong effect these points have in the integral for S^p. The $\zeta(p)$ converge to the values predicted by Eq. (29) quite rapidly with m/N. The rate of convergence is illustrated in Fig. 5 for the two cases shown in Fig. 4. Convergence is achieved at $m/N=0.001$ for $\alpha=1.8$ and $m/N=0.005$ for $\alpha=1.0$; which correspond to the largest event being $y^*\simeq 18$ and $y^*\simeq 130$, respectively. These values lie in the region given by Eq. (16), as the asymptotic tail region of the PDF is valid for $y\gg\tau^{1/\alpha}=1$ here.

It is also instructive to investigate the effects of variations in sample size N on the rates of convergence. Figure 6 illus-

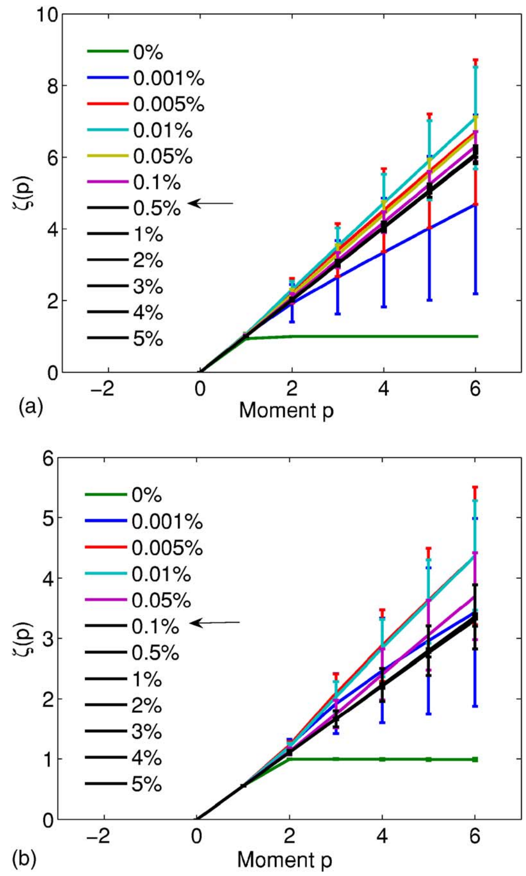

FIG. 4. (Color online) Plots showing the exponents $\zeta(p)$ against moment order p of the generalized structure functions for various values of the percentage of large events excluded for (a) $\alpha=1.0$ and (b) $\alpha=1.8$. The arrows indicate the percentage beyond which convergence to the expected behavior $\zeta(p)=p/\alpha$ is established. Both plots are for a sample size of $N=10^6$.

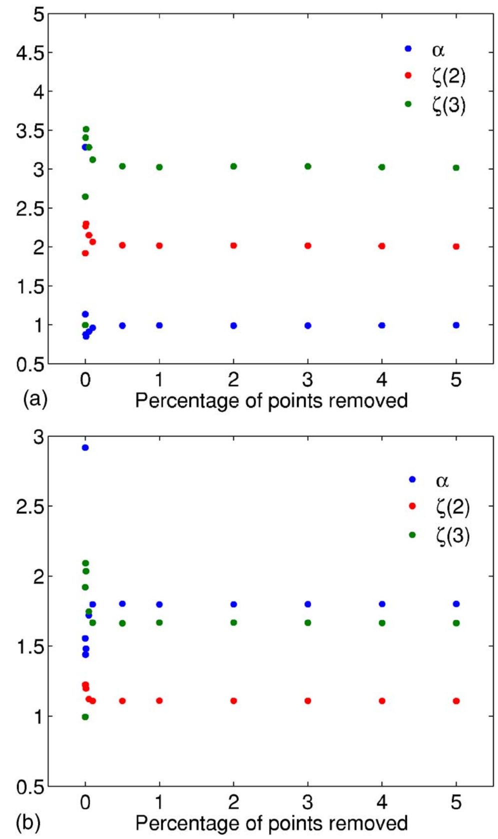

FIG. 5. (Color online) Plots showing the rapid convergence of the Lévy parameter α; and the exponents of the 2nd and 3rd moments $\zeta(2)$ and $\zeta(3)$. The plots in (a) are for $\alpha=1.0$ and in (b) for $\alpha=1.8$—both have $N=10^6$. $\zeta(2)$ and $\zeta(3)$ are the best fit gradients of the S^p vs τ plots, and α is obtained from the inverse of the gradient of the $\zeta(p)$ vs p plot shown in Fig. 4.

trates these effects in the form of $\zeta(p)$ vs p plots for sizes $N=10^5$ and $N=5\times10^6$ for a Lévy process of index $\alpha=1.0$. Recall that decreasing the sample size would result in further undersampling and thus poor statistics in the tails of the PDF. This can be clearly seen in Fig. 6(a) where we see a slow convergence to the line $\zeta(p)=p/\alpha$ which is achieved after $\sim4\%$ of the data is excluded. The converse of this is shown in Fig. 6(b) where increasing the sample size by a factor of 20 results in a very rapid convergence to scaling which is reached after only $\sim0.5\%$ of the data is excluded.

Lastly we consider the behavior of the outliers that are removed by this procedure. As we successively remove more outliers (increasing m), the behavior of y_m^* will more closely correspond to that of $\overline{y}_m^*$. This is shown in Fig. 7 where we plot $y_m^*(\tau)$ for increasing m/N. The anticipated scaling (25) appears at a value of m/N corresponding to a few percent. A more established method for determining the scaling of outliers is a rank order (or Zipf) plot (see Sornette [2]); this is shown in Fig. 8 where we plot $y_m^*(m/N)$ for successively large values of τ. The scaling with m/N is again as expected from Eqs. (20)–(24), and the rank order plots also highlight scatter of individual realizations of y_m^* from the ensemble average. In Fig. 8 this becomes apparent at higher values of τ. As we increase τ we require a higher fraction of points to be excluded before we regain the expected scaling with m/N. This breakdown of the scaling at higher values of τ follows from Eqs. (30) and (31). We can see that Λ increases with τ and so the distribution becomes more broad. Consequently this will require a higher fraction m/N of points to be excluded so that we may regain the scaling behavior (20). At the largest τ, Figs. 7 and 8 show a saturation indicative of the difference y_m^* being dominated by a single extremal value x of the original time series in Eq. (1). These plots are also a useful indicator of how feasible, for a dataset of size N, it would be to distinguish a departure from Lévy scaling in the tails.

IV. SUMMARY AND CONCLUSIONS

In this paper we have presented a technique for "conditioning" data to deal with anomalous scaling properties that arise due to finite-size effects. We have demonstrated our ideas on a numerically generated symmetric α-stable Lévy process. We are concerned with the situation of observations of natural systems, or of experiments, where the underlying PDF is not known *a priori* and where one inevitably has a finite length series of data. Hence we have proposed a technique that does not require *a priori* knowledge of the under-

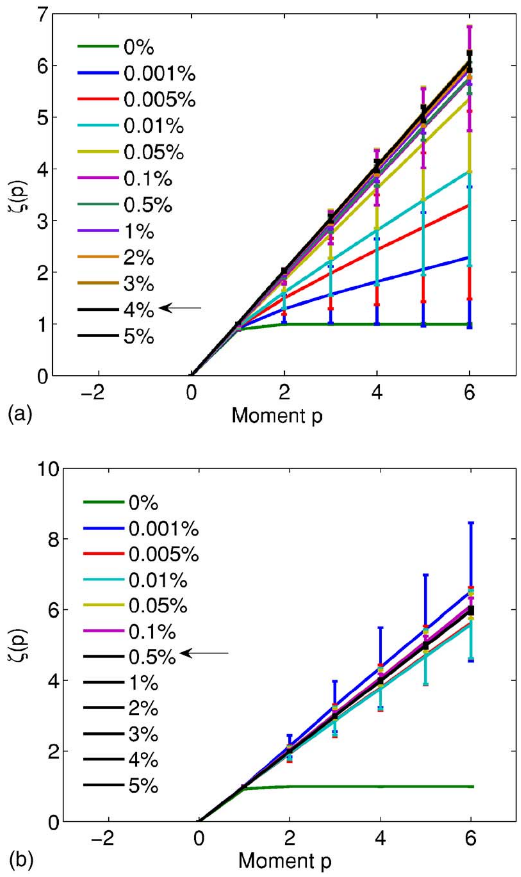

FIG. 6. (Color online) $\zeta(p)$ vs p plots for $\alpha=1.0$; (a) $N=10^5$ and (b) $N=5\times10^6$. The arrows indicate the percentage beyond which convergence to the expected behavior $\zeta(p)=p/\alpha$ is established.

lying process and that has consistency checks.

We have shown that "conditioning" the data by progressively excluding the outliers, or extremal points, when computing the scaling exponents from the structure functions, recovers the underlying scaling of a self-affine process up to large order. For large datasets of a Lévy process this corresponds to removing 0.1–1% of the data. The conditioned

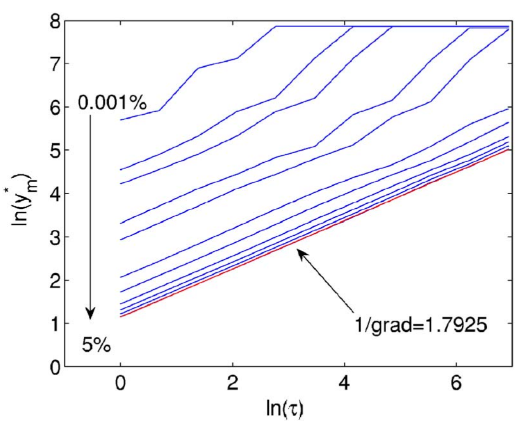

FIG. 7. (Color online) Log-log plot illustrating the scaling of the mth largest event y_m^* with τ as m is increased; $\alpha=1.8$, $N=10^6$. For comparison with previous figures we indicate the percent of points that would be excluded for the particular m.

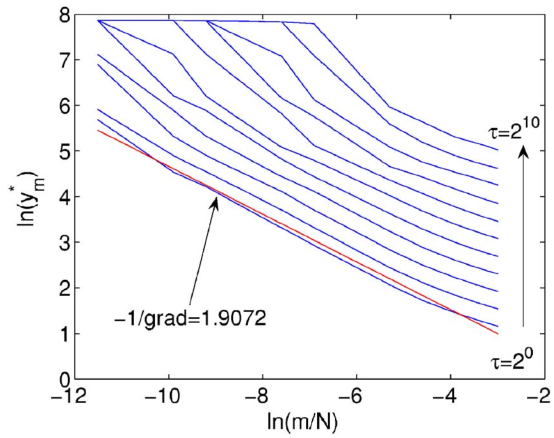

FIG. 8. (Color online) Log-log plot illustrating the scaling of the largest event y_m^* with m/N for various values of τ; $\alpha=1.8$, $N=10^6$.

structure functions then provide a straightforward method for determining the self-affine scaling exponent, in this case the Lévy index α, directly from the slope of a plot of the exponents versus moment order.

This method offers two consistency checks. The first of these is that for a self-affine process, as we progressively remove more outliers we expect that the exponents obtained from the structure functions should converge on values which then do not vary. Practically speaking, one would plot the exponents as a function of the location of the last outlier excluded and expect a plateau that extended deep into the tail of the PDF. A second check is obtained by examining the scaling properties of these discarded outliers.

Importantly, the above analysis assumes that we have some relatively good statistics—in practice the high variability of the Lévy process due to the fat tails will always result in some lone extreme points with a finite probability of occurrence, resulting in anomalous scaling exponents. This implies that we *always* need some way of cleaning or conditioning the data to recover the scaling behavior. These lone points can have a drastic effect since in a Lévy-like process the largest value of a set of increments of a time series can be of the order of the total sum [2,8]. Coupled with this we have that the tails of a distribution are described by the higher order moments (structure functions here). If the statistics of the tail are not well resolved then these moments will also give anomalous values of $\zeta(p)$.

Although we have chosen to study the simple case of monofractal scaling, the methods described in this paper hold for a general self-affine system where the scaling is described by a function $f(b)$ [see Eq. (5)]. The extension of the work presented in this paper to more general self-affine scalings, e.g., bifractal scalings; and, in principle, multiaffine time series will be the subject of further work.

ACKNOWLEDGMENTS

The authors would like to thank N. Watkins and G. Rowlands for helpful discussions and suggestions. K.K. acknowledges the financial support of the UK Particle Physics and Astronomy Research Council.

[1] J. P. Sethna, K. A. Dahmen, and C. R. Myers, Nature (London) **410**, 242 (2001).

[2] D. Sornette, *Critical Phenomena in Natural Sciences* (Springer-Verlag, Berlin, 2000).

[3] B. B. Mandelbrot, *The Fractal Geometry of Nature* (Freeman, New York, 1983).

[4] G. M. Zaslavsky, Phys. Rep. **371**, 461 (2002).

[5] F. Schmitt, D. Schertzer, and S. Lovejoy, Appl. Stochastic Mod. Data Anal. **15**, 29 (1999).

[6] G. M. Viswanathan, V. Afanasyev, S. V. Buldyrev, E. J. Murphy, P. A. Prince, and H. E. Stanley, Nature (London) **381**, 413 (2002).

[7] R. N. Mantegna and H. E. Stanley, Nature (London) **376**, 46 (1995).

[8] F. Bardou, J. Bouchaud, A. Aspect, and C. Cohen-Tannoudji, *Lévy Statistics and Laser Cooling* (Cambridge University Press, Cambridge, England, 2002).

[9] M. F. Shlesinger, B. J. West, and J. Klafter, Phys. Rev. Lett. **58**, 1100 (1987).

[10] B. Hnat, S. C. Chapman, and G. Rowlands, Phys. Rev. E **67**, 056404 (2003).

[11] B. Hnat, S. C. Chapman, and G. Rowlands, J. Geophys. Res. **110**, A08206 (2005).

[12] S. C. Chapman, B. Hnat, G. Rowlands, and N. W. Watkins, Nonlinear Processes Geophys. **12**, 767 (2005).

[13] N. P. Greis and H. S. Greenside, Phys. Rev. A **44**, 2324 (1991).

[14] T. Bohr, M. H. Jensen, G. Paladin, and A. Vulpiani, *Dynamical Systems Approach to Turbulence* (Cambridge University Press, Cambridge, England, 1998).

[15] G. Samorodnitsky and M. S. Taqqu, *Stable Non-Gaussian Random Processes* (Chapman & Hall, London, 1994).

[16] A. Janicki and A. Weron, *Simulation and Chaotic Behaviour of α-stable Stochastic Processes* (Marcel Dekker Inc, New York, 1994).

[17] W. Paul and J. Baschnagel, *Stochastic Processes; From Physics to Finance* (Springer-Verlag, Berlin, 1999).

[18] A. V. Chechkin and V. Y. Gonchar, Chaos, Solitons Fractals **11**, 2379 (2000).

[19] E. J. Gumbel, *Statistics of Extremes* (Columbia University Press, New York, 1967).

[20] E. Castillo, *Extreme Value Theory in Engineering* (Academic Press, New York, 1988).

[21] G. G. Katul, M. B. Parlange, and C. R. Chu, Phys. Fluids **6**, 2480 (1994).

[22] P. Veltri, Plasma Phys. Controlled Fusion **41**, A787 (1999).

[23] A. Mangeney, C. Salem, P. Veltri, and B. Cecconi, in ESA Report No. SP–492, 2001 (unpublished).

[24] S. Jespersen, R. Metzler, and H. C. Fogedby, Phys. Rev. E **59**, 2736 (1999).

[25] S. Siegert and R. Friedrich, Phys. Rev. E **64**, 041107 (2001).

[26] Note that the distributional equality $\overset{d}{=}$ is not needed here as $\bar{y}_m^*$ is a statistical quantity.