Uncovering the evolution of nonstationary stochastic variables: The example of asset volume-price fluctuations

Paulo Rocha,[1,2] Frank Raischel,[3] João P. Boto,[1,2] and Pedro G. Lind[4,5]

[1]*Centro de Matemática e Aplicações Fundamentais, Avenida Professor Gama Pinto 2, 1649-003 Lisboa, Portugal*

[2]*Departamento de Matemática, Faculdade de Ciências, University of Lisbon Campo Grande, Edifício C6, Piso 2, 1749-016 Lisboa, Portugal*

[3]*Center for Geophysics, IDL, University of Lisbon, 1749-016 Lisboa, Portugal*

[4]*ForWind—Center for Wind Energy Research, Institute of Physics, Carl-von-Ossietzky University of Oldenburg, DE-26111 Oldenburg, Germany*

[5]*Institut für Physik, Universität Osnabrück, Barbarastrasse 7, 49076 Osnabrück, Germany*

(Received 29 July 2015; revised manuscript received 29 December 2015; published 11 May 2016)

We present a framework for describing the evolution of stochastic observables having a nonstationary distribution of values. The framework is applied to empirical volume-prices from assets traded at the New York Stock Exchange, about which several remarks are pointed out from our analysis. Using Kullback-Leibler divergence we evaluate the best model out of four biparametric models commonly used in the context of financial data analysis. In our present data sets we conclude that the inverse Γ distribution is a good model, particularly for the distribution tail of the largest volume-price fluctuations. Extracting the time series of the corresponding parameter values we show that they evolve in time as stochastic variables themselves. For the particular case of the parameter controlling the volume-price distribution tail we are able to extract an Ornstein-Uhlenbeck equation which describes the fluctuations of the highest volume-prices observed in the data. Finally, we discuss how to bridge the gap from the stochastic evolution of the distribution parameters to the stochastic evolution of the (nonstationary) observable and put our conclusions into perspective for other applications in geophysics and biology.

DOI: 10.1103/PhysRevE.93.052122

I. INTRODUCTION AND MOTIVATION

When assessing the behavior of a complex system, such as the ones described by stochastic time series, one typically tries to uncover the nonlinear interactions and the strength of fluctuating forces by means of extracting an evolution equation from the data [1]. When the underlying value distributions of the observables are stationary, this approach is, in principle, possible [2]. However, in real systems the distributions are often nonstationary or, at least, it is not possible to ascertain how reasonable the assumption of stationarity is.

In this paper we address the evolution of nonstationary value distributions of stochastic observables and describe a framework that enables one to derive their evolution directly from measurements of empirical data recordings. We apply our framework to financial asset volume-prices, though the framework is general enough for many other systems, as we also discuss at the end. In particular, we show that volume-price distributions evolve in a nonstationary way but follow a typical functional shape, properly parameterized. By keeping track of the series of parameter values at each time step, we show that they follow a well-defined stochastic evolution equation, which helps to establish the evolution of the nonstationary distribution. It is known that even power laws may be derived from stochastic equations driven by Gaussian noise [3]. Further, we show how to use these findings to derive possible distribution tail boundaries that enable the estimation of risk measures. Finally, we put our results in perspective and propose a framework to fully describe the stochastic dynamics of a nonstationary variable under a few weak assumptions.

The choice of considering volume-price distributions, for example, is not arbitrary. There is an old Wall Street adage which says that "It takes volume to move price" [4]. This adage holds today. Indeed, if one considers volume or price separately from each other, one fails to grasp the behavior of the capital exchanged, which combines both variables. Therefore we consider here both variables combined, namely, the volume-price, which measures the total capital exchanged, providing information about the entire capital traded in the market.

Several articles have been written about stochastic volatility models [5–7] in order to attempt to characterize the dynamics of the stock price returns. Such models have emerged due to the well-established non-Gaussian character of financial time series [8]. For instance, asymptotic behavior consistent with a power-law decay can be found not only in price fluctuations but also in trading volumes [4,9]. Here, we find a strong competition or coexistence between a Gaussian model (log-normal) and heavy tails (inverse Γ). For computing accurate tail parameters, there is already an established panoply of tools [10–12] that can be used. Here we focus on a different approach to model the dynamics of the distribution tail and how it can be used for assessing the associated risk of gain and loss due to such large fluctuations.

We start in Sec. II by introducing four biparametric models that are typically used in finance to fit the empirical data. In Sec. III we investigate which models are best suited to explaining the empirical distributions, introducing one variant of the Kullback-Leibler divergence. In Sec. IV we reveal the time evolution of the nonstationary distribution of the volume-price based on a framework that enables one to extract a stochastic motion equation for the distribution parameters. This approach was used in the financial context recently [13] when accessing clustering states of the stock market [14]. In Sec. V we use our results to derive the evolution equations for the original nonstationary variable. Finally, in Sec. VI, we put our approach into perspective and discuss possible applications in other situations, before summarizing the main conclusions of this paper.

II. NONSTATIONARY MODELS FOR STOCHASTIC VARIABLES

Some of the most typical statistical models for stochastic variables in different fields, ranging from physics [15,16] and biology [17,18] to finance [19], medicine [20,21], and even sociology, among other fields [22], are biparametric. Moreover, they account for a range where a polynomial ansatz dominates and another which behaves exponentially. Four of the most used such biparametric distributions are the Γ distribution,

$$p_\Gamma(s) = \frac{s^{\phi_\Gamma - 1}}{\theta_\Gamma^{\phi_\Gamma}\Gamma[\phi_\Gamma]} \exp\left[-\frac{s}{\theta_\Gamma}\right], \tag{1}$$

the inverse Γ distribution,

$$p_{1/\Gamma}(s) = \frac{\theta_{1/\Gamma}^{\phi_{1/\Gamma}}}{\Gamma[\phi_{1/\Gamma}]} s^{-\phi_{1/\Gamma}-1} \exp\left[-\frac{\theta_{1/\Gamma}}{s}\right], \tag{2}$$

the log-normal distribution,

$$p_{\ln}(s) = \frac{1}{\sqrt{2\pi}\theta_{\ln} s} \exp\left[-\frac{(\log s - \phi_{\ln})^2}{2\theta_{\ln}^2}\right], \tag{3}$$

and the Weibull distribution,

$$p_W(s) = \frac{\phi_W}{\theta_W^{\phi_W}} s^{\phi_W - 1} \exp\left[-\left(\frac{s}{\theta_W}\right)^{\phi_W}\right]. \tag{4}$$

Next, we consider all these four distributions as candidate models for our data.

In each case one has two parameters, represented here by ϕ and θ, with a specific meaning. In the Γ distribution ϕ_Γ characterizes the left power tail and θ_Γ accounts for the decaying time on the right-hand side as indicated in Fig. 1(a).

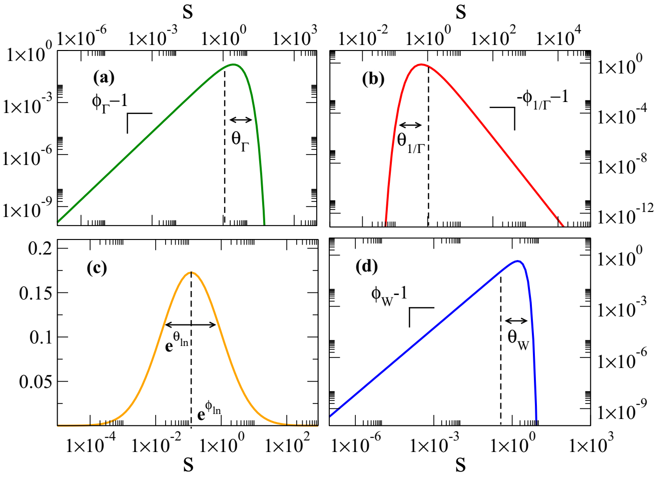

FIG. 1. The four biparametric distributions in Eqs. (1)–(4): (a) Γ distribution, (b) inverse Γ distribution, (c) log-normal distribution, and (d) Weibull distribution. For each model, a graphic illustration of its parameters is sketched.

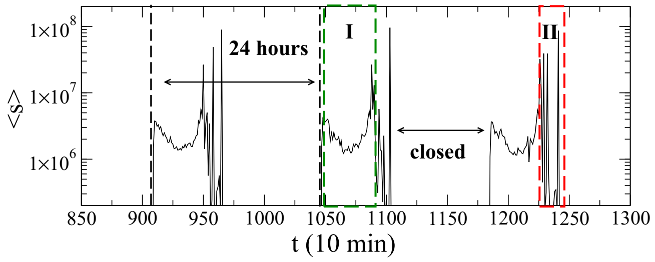

FIG. 2. Illustration of the time series of the volume-price s of one company listed in the NYSE during a period of approximately 3 days. (I) The 8-h period of normal trading; (II) afterhours trading period (after closing), which is discarded from our analysis. During the night (nontrading period) we set $s = 0$.

In the inverse Γ distribution $\phi_{1/\Gamma}$ characterizes the right power tail and $\theta_{1/\Gamma}$ accounts for the decaying time on the left-hand side of the distribution as indicated in Fig. 1(b). In the log-normal distribution $\phi_{\ln}$ accounts for the mean and $\theta_{\ln}$ for the standard deviation of the variable logarithm, as indicated in Fig. 1(c). In the Weibull distribution ϕ_W characterizes the left power tail when the exponential term goes to 1 and θ_W accounts for the decaying time on the right-hand side of the distribution as indicated in Fig. 1(d).

In the following we analyze the volume-price (s) series of around 2000 companies having listed shares on the New York Stock Exchange (NYSE), with a sampling frequency of 10 min, during a total of 976 days, which, after removing all the afterhours trading and discarding all the days with recording errors [23], contains around 1.8×10^4 data points. See the illustration in Fig. 2. All data were collected from the Web [24] and more details concerning their preprocessing may be found in Refs. [23,25,26]. Also note that in Fig. 2 it is possible to observe a U pattern, typically found in intraday volume time series [27].

In order to fit the empirical distribution of the volume-price we use the maximum likelihood scheme. The maximum likelihood scheme was applied to the probability density function above, Eqs. (1)–(4). Since in the case of stock market volume-prices the tail in the range of large values is associated with the largest fluctuations, i.e., the largest gains and losses, we concentrate our analysis on the tail of the volume-price distribution. Therefore, we use the parameters obtained from the maximum likehood scheme for the probability density functions (PDFs) to fit the corresponding empirical cumulative density function (CDF) of each one of the four models above. Figures 3(a) and 3(b) show, respectively, the probability and cumulative density function of each model (lines) that fit the empirical distribution (circles) in one particular 10-min snapshot.

When one considers the distributions in Eqs. (1) to (4) to be stationary, the parameters of each distribution are taken to be constants. In the following we introduce a different assumption: while we take a constant functional shape, i.e., one of the particular forms above, the corresponding parameters are allowed to vary in time, between two successive 10-min snapshots. In other words, we assume that for the general case of a nonstationary distribution or density function, the parameters ϕ and θ of the four models above are in fact

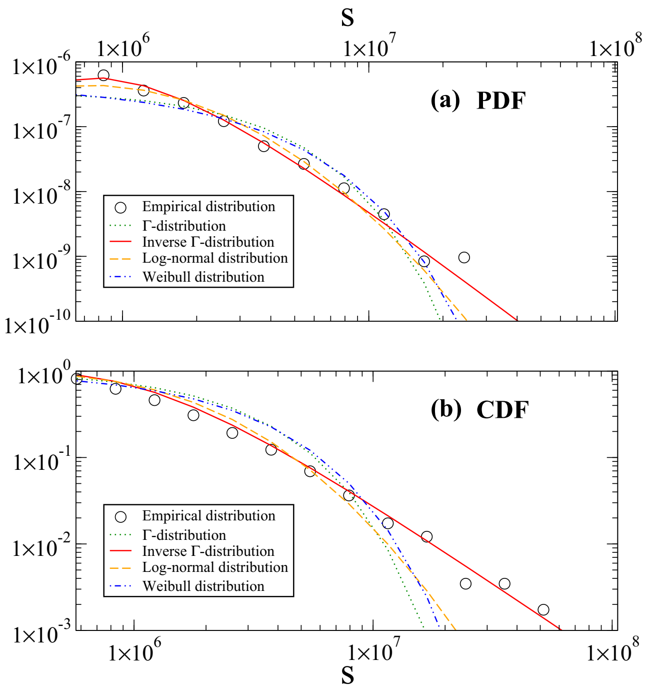

FIG. 3. Illustration of the volume-price s distribution in one particular 10-min snapshot: (a) probability density function (PDF) and (b) cumulative density function (CDF). Different colors correspond to different models used to fit the empirical data (circles).

variables of the distribution itself that include all the time dependency. In Fig. 4, we show a representation of the resulting time series of each parameter, ϕ and θ, characterizing the four models, (1)–(4), obtained from the maximum likelihood scheme. The corresponding errors $\sigma_{\hat{\theta}}$ and $\sigma_{\hat{\phi}}$ of the parameter estimators $\hat{\theta}$ and $\hat{\phi}$, respectively, are listed in Table I. Typically, the relative errors are below 10%, $0.01 \lesssim \frac{\max(\sigma_{\hat{\theta}})}{\langle\hat{\theta}\rangle}, \frac{\max(\sigma_{\hat{\phi}})}{\langle\hat{\phi}\rangle} < 0.1$. Large errors, such as the one for ϕ_Γ, indicate that the underlying distribution is probably not a good model for the empirical volume-price distribution.

TABLE I. Maximum values of the standard errors $\sigma(\hat{\phi})$ and $\sigma(\hat{\theta})$ of each estimator parameter, $\hat{\phi}$ and $\hat{\theta}$, for the different models, obtained using the maximum-likelihood estimation. Compared to the typical values of the parameters shown in Fig. 4 one can see that typical relative errors are typically between 1% and 10%.

	$\max(\sigma_{\hat{\phi}})$	$\max(\sigma_{\hat{\theta}})$
Γ distribution	0.5	3.4×10^5
Inverse Γ distribution	0.51	8.3×10^5
Log-normal	0.05	0.04
Weibull	0.07	2.3×10^5

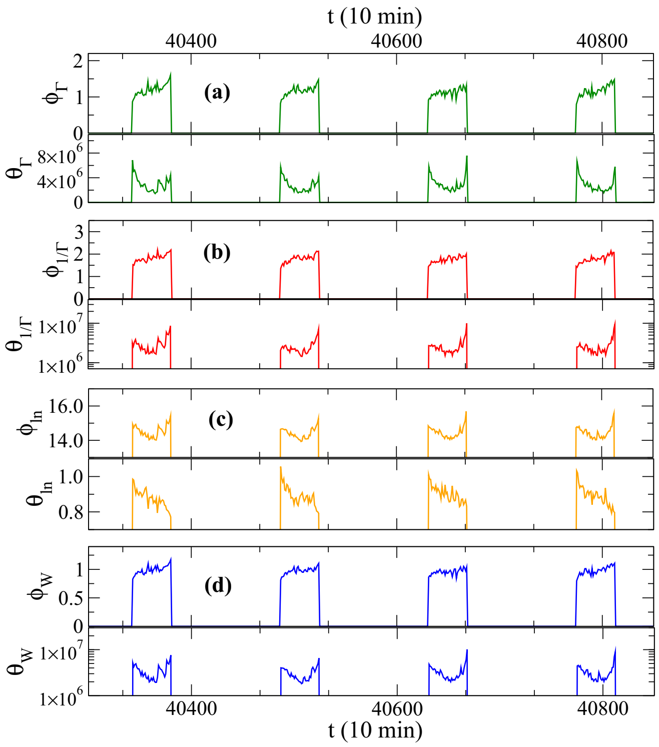

FIG. 4. Time series of the two parameters characterizing the evolution of the cumulative density function of the volume-price s: (a) Γ distribution (b) inverse Γ distribution, (c) log-normal distribution, and (d) Weibull distribution. Each point in these time series corresponds to a 10-min intervals. Periods with no activity correspond to the period when the market is closed and, therefore, is not considered in our approach. In all plots, different colors correspond to different distributions. Each parameter shows a typical daily pattern, which supports the detrending considered in Fig. 6 (see text).

III. SEARCHING FOR AN OPTIMAL MODEL OF VOLUME-PRICE DISTRIBUTIONS

In this section, we ascertain which model described previously is the best for the empirical set of volume-prices. To that end, we evaluate the accuracy of the fit of each model using a "distance" between the empirical distribution and the modeled one, which we define as

$$\begin{aligned} D^{(F)}(P|Q) &= \int_{s_{\min}}^{s_{\max}} F(s)\ln\left(\frac{P(s)}{Q(s)}\right)ds, \\ &\sim \sum_i F(s_i)\ln\left(\frac{P(s_i)}{Q(s_i)}\right)\Delta s_i, \end{aligned} \tag{5}$$

where i labels a succession of bins, covering the region of volume-price values $[s_{\min},s_{\max}]$, s_i and Δs_i represent, respectively, the mean value and the width of bin i, $P(s_i)$ is the empirical distribution, $Q(s_i)$ is the modeled PDF, and $F(s_i)$ is a weighting function. The region within which the distributions are compared is accounted for by the values of i in the sum and can cover the entire state space or a delimited subregion of it. In our case, we consider solely the tail, i.e., the values of i larger than the median of the empirical distributions.

The sum in Eq. (5) is weighted by the function $F(s_i)$. For $F(s_i) = P(s_i)$ one obtains the standard Kullback-Leibler divergence [28], where the values occurring at a higher frequency have larger weights than those occurring rarely (extreme events). Figure 5(a) shows the rankings of all four models, evaluated according to the Kullback-Leibler

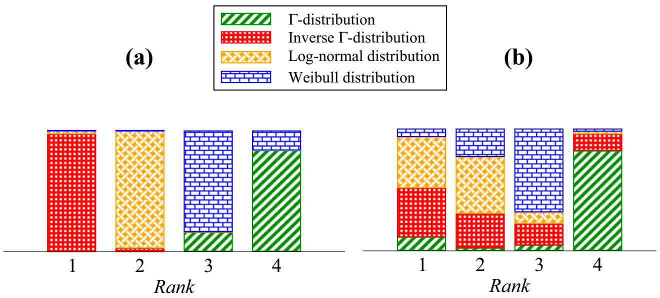

FIG. 5. What is the best model? Ranking of the four distributions in Eqs. (1)–(4) used to fit the empirical tails in two cases: (a) using the Kullback-Leibler divergence $D^{(p)}$, i.e., $F(s_i) = P(s_i)$ in Eq. (5); and (b) weighting the extreme events more strongly, with a different distance $D^{(1/p)}$ with $F(s_i) = 1/P(s_i)$. Rank 1 indicates the best model.

divergence in Eq. (5). As we can see, the best fit is almost always the inverse Γ distribution. In rank 2 one finds the log-normal distribution with only a slightly larger average distance (see Table II).

Because of the choice for the weighting function $F(s_i) = P(s_i)$, the Kullback-Leibler divergence accounts for a good fit in the central region, which is more heavily weighted than the tails. Of course, that standard case stems from information theory, where the divergence is interpreted directly from the information entropy: it is the difference between the cross entropy of P and Q and the entropy of P [29].

However, to focus on modeling the extreme events in a given empirical distribution, one needs to weight the events occurring more rarely with heavier weights, and a different weight function would be a better choice. In the following, we choose the function $F(i) = 1/P(i)$ to account for these weights.

Figure 5(b) shows the ranking for this divergence $D^{1/p}$. The results are now significantly different: the best models are the log-normal and the inverse Γ distributions. When considering the tails, there is therefore coexistence of log-normal and inverse Γ.

Table II lists the mean value and standard deviation of the value distributions of the Kullback-Leibler divergence $D^{(p)}$ and of the tail distance $D^{(1/p)}$ for all 10-min time spans. Rank 1 is dominated by the inverse Γ distribution, followed by the log-normal distribution when considering the usual Kullback-Leibler function. When the variant $D^{(1/p)}$ is chosen, one observes codominance between the log-normal and the inverse Γ distributions. Furthermore, the inverse Γ distribution is parameterized in such a way that one single parameter, $\phi_{1/\Gamma}$, controls the tail of the largest values [see Eq. (2) and Fig. 1(b)]. For all these reasons, we henceforth choose the inverse Γ distribution as our model for the evolution of the volume-price distribution tail.

TABLE II. Average and standard deviation for the Kullback-Leibler divergence $D^{(p)}$ and for the tail distance $D^{(1/p)}$. The inverse Γ model was chosen for further analysis (see text).

	$D^{(P)}$		$D^{(1/P)}$	
	Average	SD	Average	SD
Γ distribution	0.80	0.03	2.39	2.45
Inverse Γ distribution	**0.68**	**0.02**	**0.51**	**0.36**
Log-normal	0.70	0.02	**0.51**	**0.54**
Weibull	0.73	0.03	1.40	1.51

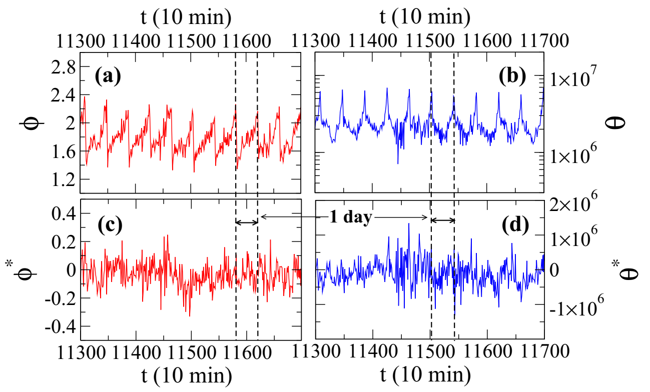

FIG. 6. Illustration of the time series of (a) ϕ (red) and (b) θ (blue), both parameters of the inverse Γ distribution in Eq. (2). (c, d) The corresponding detrended time series, ϕ^* and θ^*, are plotted (see text). In this figure we plot approximately 12 trading days, cutting out the afterhours and closing periods.

IV. STOCHASTIC EVOLUTION OF THE DISTRIBUTION TAILS

In this section we extract the stochastic evolution of the distribution tail, choosing the inverse Γ distribution as the model. For simplicity we write ϕ and θ only for the parameters $\phi_{1/\Gamma}$ and $\theta_{1/\Gamma}$.

For the analysis we first study the average time evolution of each parameter during one single day. Indeed, as we can see from Figs. 6(a) and 6(b), there is clearly a daily pattern $\bar{\phi}$ and $\bar{\theta}$, which, after being removed from the original series, yields the detrended data series of fluctuations, ϕ^* and θ^*, shown in Figs. 6(c) and 6(d), respectively. Our ansatz is therefore defined by the decomposition of the original parameter series into their daily pattern and their fluctuations:

$$\phi(t) = \bar{\phi}(t) + \phi^*(t), \tag{6a}$$

$$\theta(t) = \bar{\theta}(t) + \theta^*(t). \tag{6b}$$

Since the series is nonstationary, we consider average daily patterns for a set of 20 days. The series of fluctuations were extracted by removing the 20-day moving average pattern from the original series. This was done by centering the windows in each point of the original series and subtracting the average of the points on 10 days before and after that event.

Figure 7 shows the daily pattern of each parameter, approximated as a cubic polynomial of time,

$$\bar{\phi}(t_d) = a_\phi t_d^3 + b_\phi t_d^2 + c_\phi t_d + d_\phi, \tag{7a}$$

$$\bar{\theta}(t_d) = a_\theta t_d^3 + b_\theta t_d^2 + c_\theta t_d + d_\theta, \tag{7b}$$

where $t_d = (t \pmod{144}) - 54$ in units of $u = 10$ min. Note that the market is only open for normal trading during 6

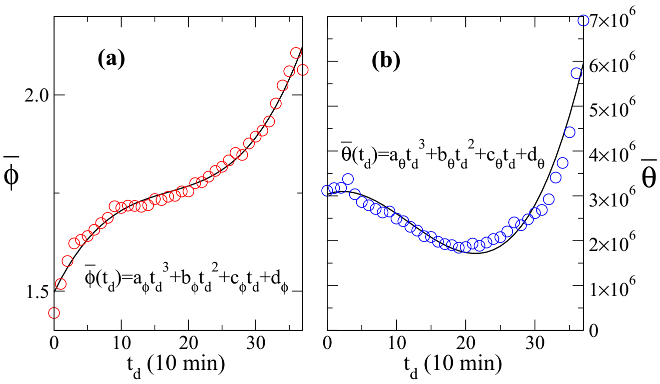

FIG. 7. Average over all trading days of parameters ϕ and θ. Here we see that both $\bar{\phi}$ and $\bar{\theta}$ seem to follow the cubic law. In the fitting function $\bar{\phi}(t_d)$ the coefficients have the values $\mathrm{a}_\phi = 3.4 \times 10^{-5}\ \mathrm{u}^{-3}$, $\mathrm{b}_\phi = -1.7 \times 10^{-3}\ \mathrm{u}^{-2}$, $\mathrm{c}_\phi = 3.5 \times 10^{-2}\ \mathrm{u}^{-1}$, and $\mathrm{d}_\phi = 1.5$. In the fitting function $\bar{\theta}(t_d)$ the coefficients have the values $\mathrm{a}_\theta = 3.9 \times 10^{2}\ \mathrm{u}^{-3}$, $\mathrm{b}_\theta = -1.4 \times 10^{4}\ \mathrm{u}^{-2}$, $\mathrm{c}_\theta = 5.7 \times 10^{4}\ \mathrm{u}^{-1}$, and $\mathrm{d}_\theta = 3.0 \times 10^{6}$ (see text), all in units of u = 10 min.

h 30 min (39 × 10 min). Outside of the normal trading period we define the $\bar{\phi}$ and $\bar{\theta}$ to be 0. From the average pattern of ϕ it is clear that large volume-prices on the NYSE tend to concentrate in the beginning of the day (low $\bar{\phi}$ values).

Figures 8(a) and 8(b) show the marginal PDFs of the variables ϕ and θ, which can be compared with the detrended variables separately [Figs. 8(c) and 8(d)]. Clearly, the detrending does not have a significant effect on the shape of the PDF of these two parameters. Figure 8(e) shows the joint PDF of ϕ^* and θ^*, from which one sees that the two parameters can be assumed to be independent of each other. Since the observed fluctuations of θ do not play a significant role in the distribution tail, we approximate parameter θ by its daily pattern, $\theta(t) \sim \bar{\theta}(t_d)$.

Under these assumptions, to fully derive the evolution equations of both parameters [Eqs. (6)] one only needs to define, additionally, the fluctuations $\phi^*(t)$, which will be modeled according to the Langevin process

$$d\phi^* = D_1(\phi^*)dt + \sqrt{2D_2(\phi^*)}dW_t, \tag{8}$$

where $D_1(\phi^*)$ and $D_2(\phi^*)$ are the so-called drift and diffusion coefficients, respectively, and dW_t is one Wiener process satisfying $\langle dW_t\rangle = 0$ and $\langle dW_t dW_t'\rangle = \delta(t-t')$.

A necessary ingredient of this approach is that ϕ^* series must be Markovian. In order to test the Markov property we compute the transition probabilities $p(x_1,\tau_1|x_2,\tau_2)$ and $p(x_1,\tau_1|x_2,\tau_2;x_3=0,\tau_3)$. In Fig. 9(a) we show the contour plot of these two probabilities for $\tau_1 = \tau_{\min}$, $\tau_2 = 5\tau_{\min}$, and $\tau_3 = 10\tau_{\min}$, with $\tau_{\min} = 10$ min. The proximity of the contour lines suggests that the Markovian property holds. Moreover, in Figs. 9(b) and 9(c), two cuts through the conditional probability densities are provided for fixed values of x_1, namely, at $\langle x_1\rangle \pm 0.4$ standard deviations of the one-point distribution $p(x_1)$, which also seems to support this statement.

In order to create a quantitative understanding of whether or not the two conditional probabilities $p(x_1,\tau_1|x_2,\tau_2)$ and $p(x_1,\tau_1|x_2,\tau_2;x_3=0,\tau_3)$ are equal, the Wilcoxon rank-sum test [30] is employed. The value of t value/t_0 value = 1 indicates that the process is Markovian. As we can see in Fig. 10, this test seems to further confirm that a proper Markov length $\tau_M = 50$ min can be reasonably assumed.

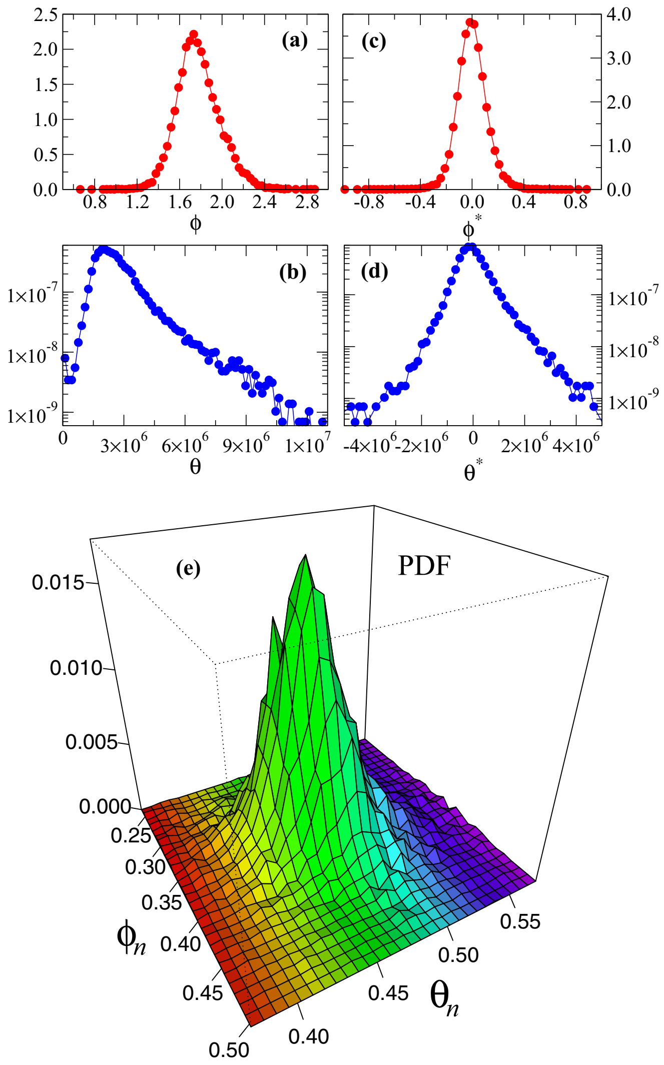

FIG. 8. Probability density function (PDF) of the fitting parameters, (a) ϕ and (b) θ, before the detrending, compared to the PDFs of their fluctuations, (c) ϕ^* and (d) θ^*, after detrending (see text). (e) The joint PDF of both detrended variables after proper normalization, $\phi_n = (\phi^* - \phi_{\min})/(\phi_{\max} - \phi_{\min})$ and $\theta_n = (\theta^* - \theta_{\min})/(\theta_{\max} - \theta_{\min})$: both detrended variables can be taken as independent of one another (see text).

For a Markovian stochastic process, the evolution of the associated stochastic variable is defined by the two functions in Eq. (8), namely, D_1 and D_2, given by

$$D_k(\phi^*) = \lim_{\tau\to 0}\frac{M_k(\phi^*,\tau)}{k!\tau} \sim \frac{M_k(\phi^*,\tau_l)}{k!\tau_l} \tag{9}$$

for $k = 1,2$ and where the conditional moments $M_k(\phi^*,\tau)$ are defined as

$$M_k(\phi^*,\tau) = \langle (X_{t+\tau} - X_t)^k\rangle_{X_t=\phi^*}. \tag{10}$$

In Figs. 11(a) and 11(b) we represent the conditional moments M_1 and M_2, respectively, as a function of τ. Computing the slopes of M_1 and M_2 for each bin in variable

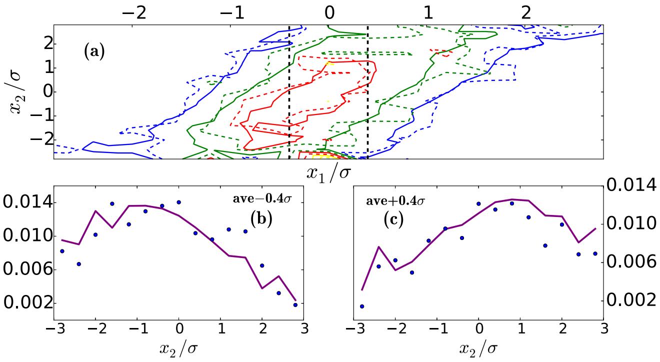

FIG. 9. (a) Contour plots of the conditional PDF $p(x_1,\tau_1|x_2,\tau_2)$ (solid lines) and $p(x_1,\tau_1|x_2,\tau_2;x_3=0,\tau_3)$ (dashed lines) for $\tau_1=\tau_{\min}$, $\tau_2=5\tau_{\min}$, and $\tau_3=10\tau_{\min}$, with $\tau_{\min}=10$ min. Dashed vertical lines at standard deviations of $\langle x_1\rangle = 0.4$ and $\langle x_1\rangle=-0.4$ indicate the cuts shown in (b) and (c), respectively.

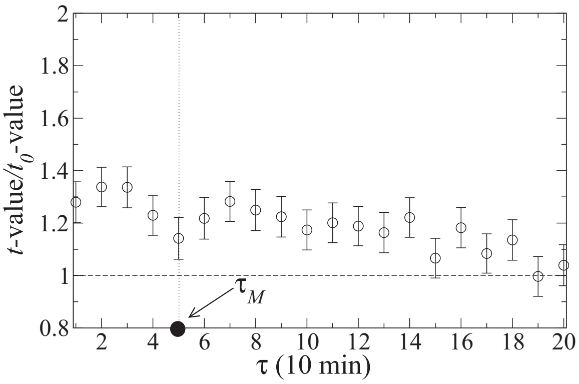

FIG. 10. Wilcoxon test to verify the Markovian property of the ϕ^* time series, showing the Markov length of $\tau_M = 50$ min.

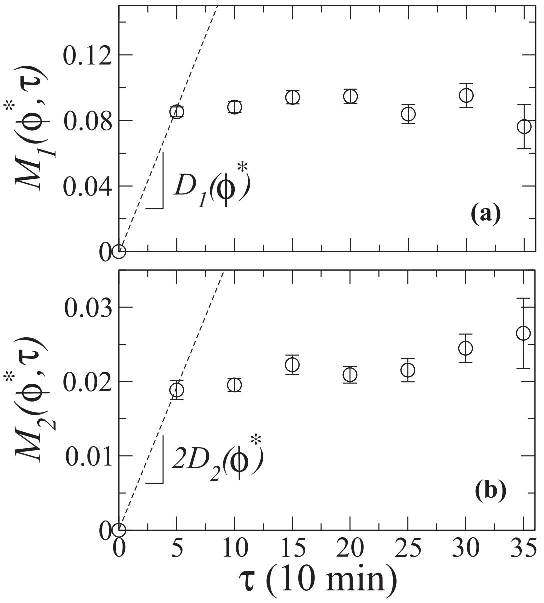

FIG. 11. Conditional moments extracted from the time series of ϕ^*. (a) First conditional moment M_1 and (b) second conditional moment M_2, both as functions of τ in units of 10 min.

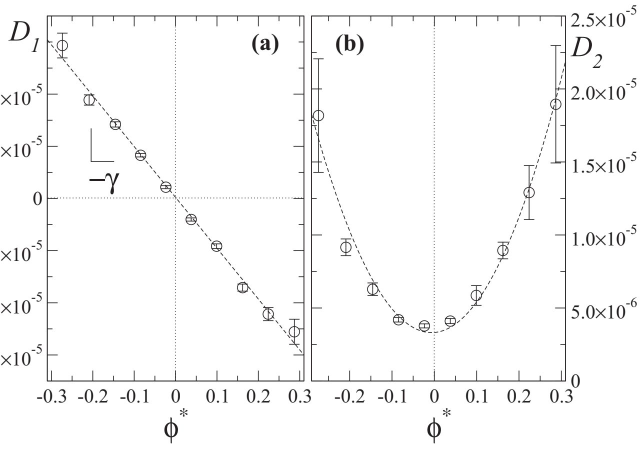

FIG. 12. Here we see that (a) the drift coefficient is linear in ϕ^*, while (b) the diffusion coefficient is a quadratic function of the stochastic variable ϕ^*. See Eqs. (11). These two coefficients characterize the stochastic evolution of the parameter ϕ, which describes the tail of the inverse Γ distribution.

ϕ yields a complete definition of both the drift D_1 and the diffusion D_2 coefficients for the full range of observed ϕ^* values.

Figures 12(a) and 12(b) show the drift and diffusion, respectively. The drift is linear on ϕ^* with a negative slope, while the diffusion is a quadratic polynomial of ϕ^*,

$$D_1(\phi^*) = -\gamma\phi^*, \tag{11a}$$

$$D_2(\phi^*) = \alpha(\phi^*)^2 + \beta\phi^* + \delta, \tag{11b}$$

with $\gamma = 1.9\times10^{-4}$ s^{-1}, $\alpha = 1.9\times10^{-4}$ s^{-1}, $\beta = 2.4\times10^{-6}$ s^{-1}, and $\delta = 3.3\times10^{-6}$ s^{-1}. Therefore the evolution of ϕ^* follows Eq. (8) with the functions as defined in (11).

V. ACCESSING THE EVOLUTION OF THE NONSTATIONARY VOLUME-PRICE

To end this paper, we present two possible applications of our framework. The first is specifically aimed for the dynamics of volume-price tails, where we provide a simple quantitative measure of the expected bounded values of the tail parameters also discussed in other contexts [31]. The second is to suggest a broader perspective, developing a framework that provides the full statistics of a nonstationary variable, based on the dynamical model of all its moments.

The first remark deals with the evolution of the original (nondetrended) ϕ parameter, following the assumption that the inverse Γ distribution is the best model for the tail at the highest volume-prices. Indeed, from the results shown in Fig. 12 and Eqs. (11) the evolution of the fluctuations ϕ^* is governed by

$$d\phi^* = -\gamma\phi^* dt + \sqrt{2(\alpha(\phi^*)^2+\beta\phi^*+\delta)}dW_t, \tag{12}$$

where $\gamma = 1.9\times10^{-4}$ s^{-1} is the inverse response time from the market to perturbations in the largest range of fluctuations. Note that γ corresponds to a response time $1/\gamma = 5.3\times10^3$ s, i.e., about 1 h 30 min, a value of the same order as the Markov length τ_M calculated in the previous section: The

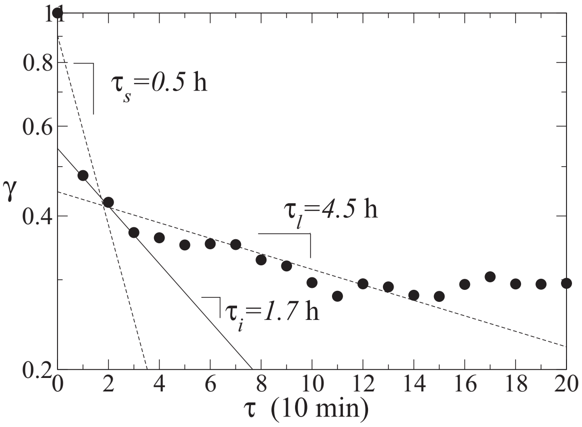

FIG. 13. Autocorrelation of ϕ^* as a function of the delay τ. A short-term and a long-term regime exist, with autocorrelation times of $\tau_s \simeq 0.5$ h and $\tau_l \simeq 4.5$ h, respectively, with dotted lines indicating corresponding exponential fits. At the time scale of our Markov modeling, the intermediate scale $\tau_i = 1/\gamma \simeq 1.7$ h, compatible with the time scale of the stochastic process modeled by Eq. (13), with the solid line representing its time scale.

market responds to perturbations at a time scale close to the time scale at which the parameter ϕ experiences stochastic variations.

As Fig. 13 indicates, the autocorrelation of ϕ^* does not follow a simple exponential decay but presents two distinct short-term and long-term regimes, one at ~30 min, associated with the trading operations, and another at ~4.5 h, which completes approximately 1 trading day. The Markov analysis captures an intermediate regime, reflecting a mixture of both time scales simultaneously and avoiding the non-Markovian properties of the short-time scale, which has an autocorrelation time of $\approx 1/\gamma$, compatible with the stochastic process in Eq. (12).

According to Eq. (6a) we can now write the evolution equation for the tail parameter ϕ, according to $d\phi = d\bar{\phi} + d\phi^*$, which, from Eq. (12), reads

$$d\phi = -\gamma(\phi - \phi_f)dt + \sqrt{2\tilde{D}_2(\phi)}dW_t, \tag{13}$$

where $\tilde{D}_2(\phi)$ is also a quadratic function of the argument ϕ, of course with different coefficients, and ϕ_f is a fixed point depending only on the average tail slope $\bar{\phi}$:

$$\phi_f = \bar{\phi} + \frac{1}{\gamma}\frac{d\bar{\phi}}{dt}. \tag{14}$$

From Eq. (14) it is clear that ϕ_f depends exclusively on the time of day, t_d, since it is completely defined by the daily pattern $\bar{\phi}$. From the coefficients of $\bar{\phi}$ [see Fig. 7 and Eq. (14)] one sees that the drift term is larger at the beginning and end of each trading-day. At the market opening, the strong drift reflects the strong herd reaction from the market accumulated during the night, pushing the distribution tail towards the median value of $\bar{\phi}$. As stated above, the cumulative buy and sell demands during the night also explain the lower average tail slope, i.e., the higher volume-prices observed in the corresponding distribution tail. At closing, a stronger drift occurs, again reflecting the pressure of buyers and sellers to match the present state and tendency of the market before closing. Note that all together the drift term in Eq. (14) is typically positive only because of the average pattern, whose dynamics is ruled by the large term in ϕ_f, namely, $\frac{1}{\gamma}\frac{d\bar{\phi}}{dt}$. See Fig. 14.

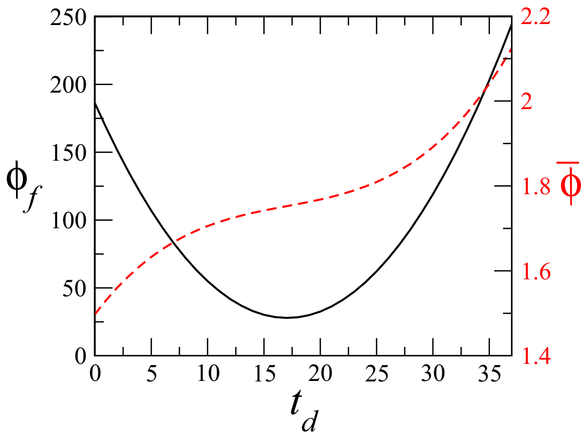

FIG. 14. For the NYSE, while the average tail slope, $\bar{\phi}$, varies cubically within 1 day, the corresponding fixed point ϕ_f [see Eq. (14)] varies mostly quadratically, since its cubic coefficient is vary small compared to the other terms in the polynomial. Such findings lead to insight into the average volume-price behavior on the NYSE.

It is also important to note that the standard deviation σ of the distribution of observed ϕ values can be estimated from and compared with

$$\sigma^2 = \int_0^\infty (\phi - \phi_f)^2 P_{\rm stat}(\phi) d\phi, \tag{15}$$

where $P_{\rm stat}(\phi)$ is the stationary distribution of ϕ, in the sense that there exists a stationary distribution when averaged over multiples of the 1-day periods, for given drift and diffusion functions. In our approach drift and diffusion are defined in Eq. (13), with ϕ given in Eq. (12), the average $\bar{\phi}$ defined by the cubic polynomial in Eq. (7a), and the fluctuation ϕ^* governed by drift and diffusion in Eqs. (11).

As shown in the Appendix, the integral in Eq. (15) exists if

$$\frac{\gamma}{\alpha} > 1, \tag{16}$$

which is intuitive: the standard deviation of the ϕ distribution only exists when the drift is strong enough, i.e., when the slope $-\gamma$ [see Fig. 12(a)] is steep enough to dominate the diffusion ($\gamma > \alpha$). For the NYSE this is not the case: $\alpha = \gamma$ within numerical accuracy. Therefore, we may claim that for the highest range of volume-prices on the NYSE, an estimate of the risk associated with the predictions of the distribution tail is doubtful.

The second remark deals with the description of the (nonstationary) evolution of the original stochastic variable, in this case the volume-price s. As we have seen, while for the large-value tail the inverse Γ model yields a good and simple description of its evolution, the log-normal distribution is found to be a good model for the tail as well. In this case, both parameters of the distribution in Eq. (3) must be considered together. In general, the distribution changes with time, due to the fluctuations of ϕ and θ, but still we can assume that all time dependency is incorporated in the distribution parameters, $P_{\ln}(s,t) \equiv P(s,\phi(t),\theta(t))$.

If one is able to write all the moments of s as a function of the distribution parameters $\phi(t)$ and $\theta(t)$, one is able to fully characterize the nonstationary evolution of s. Indeed, the moments of s can be generally written as

$$\langle s^n \rangle = \int_0^{+\infty} s^n P(s,\phi(t),\theta(t)) ds \equiv F_n(\phi(t),\theta(t)) \tag{17}$$

when the integral exists. In the most general case, both parameters can be taken as stochastic variables coupled to each other and, therefore, obeying the Langevin system of equations [2,32],

$$d\phi = h_1(\phi,\theta)dt + g_{11}(\phi,\theta)dW_1 + g_{12}(\phi,\theta)dW_2, \tag{18a}$$

$$d\theta = h_2(\phi,\theta)dt + g_{21}(\phi,\theta)dW_1 + g_{22}(\phi,\theta)dW_2, \tag{18b}$$

where $(h_1,h_2)^{(T)} = D_1$ and $gg^T = D_2$. Deriving function F in Eq. (17) one extracts the evolution equation of all statistical moments by differentiating Eq. (17) using Itô-Taylor expansion and incorporating the Eqs. (18), namely [32],

$$d\langle s^n \rangle = A_n(\phi(t),\theta(t))dt + B_n(\phi(t),\theta(t))dW_1 + C_n(\phi(t),\theta(t))dW_2, \tag{19}$$

with

$$A_n(\phi(t),\theta(t)) = \frac{\partial F_n}{\partial \phi}h_1 + \frac{\partial F_n}{\partial \theta}h_2 + \frac{\partial^2 F_n}{\partial \phi \partial \theta}(g_{11}g_{21} + g_{12}g_{22}) + \frac{1}{2}\frac{\partial^2 F_n}{\partial \phi^2}(g_{11}^2 + g_{12}^2) + \frac{1}{2}\frac{\partial^2 F_n}{\partial \theta^2}(g_{21}^2 + g_{22}^2), \tag{20a}$$

$$B_n(\phi(t),\theta(t)) = \frac{\partial F_n}{\partial \phi}g_{11} + \frac{\partial F_n}{\partial \theta}g_{21}, \tag{20b}$$

$$C_n(\phi(t),\theta(t)) = \frac{\partial F_n}{\partial \phi}g_{12} + \frac{\partial F_n}{\partial \theta}g_{22}. \tag{20c}$$

Equation (19) is a nonhomogeneous stochastic differential equation with "drift" and "diffusion" functions which depend on time.

VI. DISCUSSION AND CONCLUSIONS

In this paper we study the stochastic evolution of the volume-price distributions of assets traded on the New York Stock Exchange as a prototypical example of nonstationary distributions of stochastic variables. We have shown that these distributions are nonstationary, in the sense that the parameters characterizing the distribution are themselves stochastic variables. In order to find the best fit for the volume-price distribution we tested four biparametric models commonly used in modeling the price of financial assets [19], namely, the Γ distribution, inverse Γ distribution, log-normal distribution, and Weibull distribution.

To weight each value in the volume-price spectrum according to some density function we use the Kullback-Leibler divergence and one new variant, Eq. (5), which accounts for extreme events, and present evidence that the inverse Γ distribution is at least a very reasonable choice for modeling the region of the spectrum of highest values. Moreover, attending to the fact that in the inverse Γ distribution the two parameters decouple, we focus our study on the parameter ϕ, which characterizes the large fluctuations of the volume-price distribution. By applying the framework in Ref. [2], we are able to extract a stochastic differential equation that describes the evolution of this parameter, Eq. (13). In general, taken together with previous investigations, e.g., agent-based models for studying transitions between different regimes of trading in a financial network [33], this will eventually permit the derivation of risk measures for the largest fluctuations. However, for the specific case of the NYSE, we have provided evidence that diffusion of the volume-price tail is too large to enable bounded values of its slope fluctuations. Thus, risk estimates associated with the tail distribution are, at most, difficult and perhaps doubtful or not possible.

We have also provided a framework for deriving the stochastic evolution of a nonstationary variable, under the assumption that it follows a biparametric model whose parameters are themselves stochastic variables in time incorporating all the time dependency of the nonstationary process. By computing all the moments as a function of these distribution parameters, one is able to fully characterize the nonstationary evolution of the stochastic variable. In particular, this approach may be helpful in other situations and applications, such as in biology, when accessing the evolution of heart interbeat intervals, or in energy sciences, to address nonstationary measurement series in energy power production of wind turbines.

Finally, the tail alone of one distribution can alternatively be modeled through a generalized Pareto distribution (GPD), which yields three parameters (location, shape, and scale). While in this work we have shown a model for the fundamental dynamics of volume-price distribution tails using one single parameter, in a step forward approach, the generalized Pareto distribution could be used, yielding a three-dimensional system of coupled stochastic equations for the evolution of all three parameters. While such an approach increases the complexity of our approach considerably, which reduces the fundamental dynamics of the distribution tail to the stochastic modeling of one single parameter, new insight could be extract from this more generalized model and add criticism to the common models used for volume-price distribution and dynamics. These issues will also be considered in forthcoming studies.

ACKNOWLEDGMENTS

The authors thank Philip Rinn, Matthias Wächter, and David Bastine for useful discussions and for provision of the source code for the Wilcoxon test and the Langevin analysis. P.R. thanks Fundação para a Ciência e a Tecnologia and Centro de Matemática e Aplicações Fundamentais for financial support during the time of development of this project and Deutscher Akademischer Austauschdienst for the intership fellowship through IPID4all from University of Oldenburg. P.R. and J.P.B. thank Faculdade de Ciências da Universidade de Lisboa for providing working accommodation. F.R. thanks Fundação para a Ciência e a Tecnologia (FCT) for fellowship SFRH/BPD/65427/2009. P.G.L. thanks

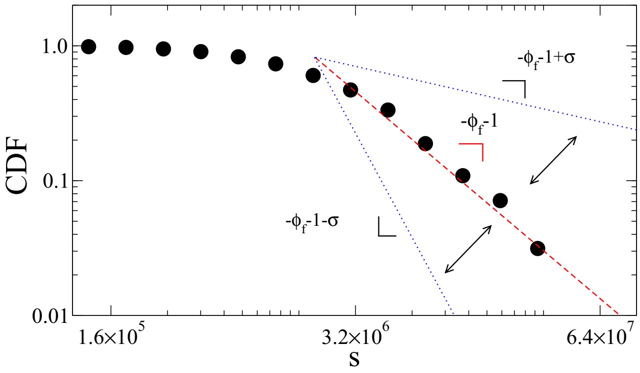

FIG. 15. We have an harmonic restoring force mechanism due to the linear drift coefficient in the stochastic differential equation characterizing the evolution of the parameter ϕ.

the German Environment Ministry for financial support. P.G.L. and F.R. thank Deutscher Akademischer Austauschdienst (DAAD) and FCT for support from bilateral collaboration DRI/DAAD/1208/2013.

APPENDIX: IS THERE A RISK ESTIMATE FOR THE TAIL OF NYSE VOLUME-PRICE DISTRIBUTIONS?

As sketched in Fig. 15, having ϕ_f in Eq. (14) and σ in Eq. (15), one can establish the upper and lower bounds for the tail of the empirical distribution, namely, $-\phi_f - \sigma$ and $-\phi_f + \sigma$, respectively, which may be helpful to derive risk measures for large fluctuations. In this Appendix, we show that, for our NYSE data, no standard deviation exists.

We, first, write the drift and diffusion functions in Eq. (13), which follow directly from substituting $\phi^* = \phi - \bar{\phi}$ in Eqs. (11), yielding

$$\tilde{D}_1(\phi,t_d) = -\gamma(\phi - \phi_f), \tag{A1a}$$

$$\begin{aligned}\tilde{D}_2(\phi,t_d) &= \alpha(\phi - \bar{\phi})^2 + \beta(\phi - \bar{\phi}) + \delta \\ &= \tilde{\alpha}\phi^2 + \tilde{\beta}\phi + \tilde{\delta},\end{aligned} \tag{A1b}$$

with

$$\tilde{\alpha} = \alpha, \tag{A2a}$$

$$\tilde{\beta} = \beta - 2\alpha\bar{\phi}, \tag{A2b}$$

$$\tilde{\delta} = \alpha\bar{\phi}^2 - \beta\bar{\phi} + \delta, \tag{A2c}$$

all coefficients now depending on the time of day t_d.

It is known [1] that the stationary probability density function $P_{\text{stat}}(\phi)$ of a nonlinear Langevin process $\phi(t)$, governed by drift $D_1(\phi)$ and diffusion $D_2(\phi)$, is given by

$$P_{\text{stat}}(\phi) = \frac{N}{D_2(\phi)} \exp\left(\int_\phi \frac{D_1(\phi')}{D_2(\phi')} d\phi'\right), \tag{A3}$$

where N is a normalization constant. Thus, substituting Eqs. (A1) into (A3) yields

$$\begin{aligned}P_{\text{stat}}(\phi) &= N(\tilde{\alpha}\phi^2 + \tilde{\beta}\phi + \tilde{\delta})^{-\frac{\gamma}{2\tilde{\alpha}}-1} \exp \\ &\times \left(\frac{2\phi_f}{\sqrt{\tilde{\alpha}\tilde{\delta} - \tilde{\beta}^2}} \arctan\left(\frac{2\tilde{\alpha}\phi_f + \tilde{\beta}}{\sqrt{\tilde{\alpha}\tilde{\delta} - \tilde{\beta}^2}}\right)\right).\end{aligned} \tag{A4}$$

Note that the radicals in the expression above always exist, since the randicands are positive, i.e., $\tilde{D}_2$ always has a negative discriminant.

Finally, substituting Eq. (A4) into Eq. (15) yields

$$\begin{aligned}\sigma^2 = N\int_0^\infty &\frac{(\phi - \phi_f)^2}{(\tilde{\alpha}\phi^2 + \tilde{\beta}\phi + \tilde{\delta})^{\frac{\gamma}{2\tilde{\alpha}}+1}} \exp \\ &\times \left(\frac{2\phi_f}{\sqrt{\tilde{\alpha}\tilde{\delta} - \tilde{\beta}^2}} \arctan\left(\frac{2\tilde{\alpha}\phi_f + \tilde{\beta}}{\sqrt{\tilde{\alpha}\tilde{\delta} - \tilde{\beta}^2}}\right)\right) d\phi.\end{aligned} \tag{A5}$$

Clearly, the exponential function is limited in the range of integration. Thus, the integral exists and σ has a finite value if the difference in the degrees of each polynomial is larger than 1, $2(\frac{\gamma}{2\tilde{\alpha}} + 1) - 2 > 1$, yielding the condition in Eq. (16).

While for the specific case of the NYSE there is a diverging standard deviation, it may be the case that similar stochastic processes governing the distribution tail of volume-prices in other stock markets have a stronger drift force or a weaker diffusion, enabling us to estimate bounding values for the slope of the distribution tail as sketched in Fig. 15.

[1] H. Risken, *Fokker-Planck Equation* (Springer, Berlin, 1984).

[2] R. Friedrich, J. Peinke, M. Sahimi, and M. Tabar, Phys. Rep. **506**, 87 (2011).

[3] J. Ruseckas and B. Kaulakys, Phys. Rev. E **81**, 031105 (2010).

[4] P. Gopikrishnan, V. Plerou, X. Gabaix, and H. E. Stanley, Phys. Rev. E **62** R4493(R) (2000).

[5] D. Delpini and G. Bormetti, Phys. Rev. E **83**, 041111 (2011).

[6] J. F. Muzy, R. Baïle, and E. Bacry, Phys. Rev. E **87**, 042813 (2013).

[7] M. Zamparo, F. Baldovin, M. Caraglio, and A. L. Stella, Phys. Rev. E **88**, 062808 (2013).

[8] A. Gerig, J. Vicente, and M. A. Fuentes, Phys. Rev. E **80**, 065102(R) (2009).

[9] X. Gabaix, P. Gopikrishnan, V. Plerou, and H. Stanley, Nature **423**, 267 (2003).

[10] A. McNeil, R. Frey, and P. Embrechts, *Quantitative Risk Management: Concepts, Techniques and Tools* (Princeton University Press, Princeton, NJ, 2015).

[11] S. Coles, *An Introduction to Statistical Modeling of Extreme Values* (Springer, Berlin, 2001).

[12] A. Clauset, C. Shalizi, and M. Newman, SIAM Rev. **51**, 661 (2009).

[13] P. Rinn, Y. Stepanov, J. Peinke, T. Guhr, and R. Schäfer, Europhys. Lett. **110**, 68003 (2015).

[14] M. C. Münnix, T. Shimada, R. Schäfer, F. Leyvraz, T. H. Seligman, T. Guhr, and H. E. Stanley, Sci. Rep. **2**, 644 (2012).

[15] J. H. Lienhard and P. L. Meyer, Ann. Math. Stat. **25**, 330 (1967).

[16] I. Eliazar, Phys. Rev. E **86**, 031103 (2012).

[17] P. Comtois, Aerobiologia **16**, 171 (2000).

[18] X. Xia, *Data Analysis in Molecular Biology and Evolution, Vol. 1* (Kluwer Academic, Dordrecht, Netherlands, 2002), p. 254.

[19] S. Camargo, S. M. D. Queirós, and C. Anteneodo, Eur. Phys. J. B **86**, 159 (2013).

[20] H. P. Zhu, X. Xia, C. H. Yu, A. Adnan, S. F. Liu, and Y. K. Du, BMC Gastroenterol. **11** (2011).

[21] H. Shen, L. Brown, and H. Zhi, Stat. Med. **25**, 3023 (2006).

[22] E. Limpert, W. A. Stahel, and M. Abbt, BioScience **51**, 341 (2001).

[23] P. Rocha, F. Raischel, J. Cruz, and P. G. Lind, in *The 3rd Stochastic Modeling Techniques and Data Analysis International Conference (SMTDA2014), June 11–14, Lisbon Portugal* (2015), pp. 619–627.

[24] http://finance.yahoo.com/.

[25] P. Rocha, F. Raischel, J. Boto, and P. G. Lind, J. Phys.: Conf. Ser. **574**, 012148 (2014).

[26] P. Rocha, Master's thesis, University of Lisbon, 2014.

[27] A. R. Admati and P. Pleiderer, Rev. Financ. Stud. **1**, 3 (1988).

[28] S. Kullback and R. Leibler, Ann. Math. Stat. **22**, 79 (1951).

[29] A. Lesne, Math. Struct. Comp. Sci. **24**, e240311 (2014).

[30] F. Wilcoxon, Biometr. Bull. **1**, 80 (1945).

[31] J. P. da Cruz and P. G. Lind, Phys. Lett. A **377**, 189 (2013).

[32] V. V. Vasconcelos, F. Raischel, M. Haase, J. Peinke, M. Wächter, P. G. Lind, and D. Kleinhans, Phys. Rev. E **84**, 031103 (2011).

[33] S. Cantono and S. Solomon, New J. Phys. **12**, 075038 (2010).