Effectiveness of the Kozachenko-Leonenko estimator for generalized entropic forms

Sílvio M. Duarte Queirós*
Unilever R&D Port Sunlight, Quarry Road East, Wirral CH63 3JW, United Kingdom
(Received 19 June 2009; published 3 December 2009)

In this Brief Report we discuss the effectiveness of the Kozachenko-Leonenko entropy estimator when generalized to cope with entropic forms customarily applied to study systems evincing asymptotic scale invariance and dependence (either of linear or nonlinear kind). We show that when the variables are independently and identically distributed the estimator is only valuable along the whole domain if the data follow the uniform distribution, whereas for other distributions the estimator is only effectual in the limit of the Boltzmann-Gibbs-Shanon entropic form. We also analyze the influence of the dependence (linear and nonlinear) between variables on the accuracy of the estimator between variables. As expected in the last case the estimator loses efficiency for the Boltzmann-Gibbs-Shanon entropic form as well.

DOI: 10.1103/PhysRevE.80.062101 PACS number(s): 05.70.−a, 02.70.Rr, 05.45.Tp, 02.60.Gf

I. INTRODUCTION

After a period of debate, the connection between the microscopic world and the displayed macroscopic properties of the system by means of the Boltzmann principle, which was later extended by Gibbs to systems in contact with a reservoir, has achieved an incontestable consensus [1]. Despite its broad acceptance, it is neglected by many people that the standard statistical mechanics is still based on a hypothesis, the *Stosszahl Ansatz* [2]. This ansatz is intimately related to the ergodic theory which has only been analytically proven for a set of very few simple systems [3]. With the surging interest in more intricate systems for which the ergodic theory is bound to be invalid, e.g., systems that occupy their allowed phase space in a scale-invariant way or exhibit long spatiotemporal correlations [4], entropic forms different to the Boltzmann-Gibbs (BG) functional have been presented. Among several, two of them might be given special emphasis: the Renyi entropy [5] and the nonadditive entropy proposed in a physical context by Tsallis [6]. For the last two decades there has been an impressive amount of work toward the physical validation and application of the latter [8]. As occurs in the BG standard case [9–11], many systems studied within the nonadditive formalism present a reduced number of observations or correspond to finite-size systems [12,13]. Consequently, a considerable error can be introduced if the simplest method based on binning the data is assumed and the number of observables is very small.

In this Brief Report we generalize a well-known binless strategy for the estimation of BG entropy, the Kozachenko-Leonenko algorithm (KLA) [14], which bases the estimation of the theoretical entropy on the distance $\delta/2$ to the nearest neighbor of a specific order n. We illustrate its possible validity by comparing numerical results with the theoretical values in two different situations: independent and dependent variables. In the former, we survey three standard distributions (PDF), namely, the Gaussian, the Student-t (or q-Gaussian), and the uniform PDF. In the latter case, we analyze linear and nonlinear dependent Student-t variables. For the sake of simplicity, we will restrict our analysis to one-dimensional systems corresponding to sets of random variables.

II. GENERALIZING KLA

The nonadditive entropy is defined as [6]

$$S_Q \equiv \frac{1-\int [p(x)]^Q dx}{Q-1} \quad (Q \in \mathbb{R}), \tag{1}$$

which in the limit Q going to 1 concurs with the Boltzmann-Gibbs entropy, $S_1=S_{BG}\equiv-\int p(x)\ln p(x)dx=-\langle \ln p(x)\rangle$, where $\langle\cdots\rangle$ represents the average. Bearing in mind the Q-logarithm definition [8], $\lim_{Q\to 1}\{\ln_Q x\equiv (x^{1-Q}-1)/(1-Q)\}=\ln x$, it is easily verifiable that the entropic functional can be written in the following way:

$$S_Q^* = -\int p(x)\ln_q p(x)dx = -\langle \ln_q p(x)\rangle, \tag{2}$$

with $Q=2-q$. In other words, the entropy S_Q represents the average value of an alternative way of describing the *surprise*. From this definition, we are evoked to apply the same ideas of the binless KLA.

Let us consider a set of N random variables, $\{x_i\}$, identically distributed and associated with a generic PDF, $p(x)$, whose entropy estimation works out at

$$S_Q = -\frac{1}{N}\sum_i \ln_q P(x_i) \equiv -\langle \ln_q P_i\rangle, \tag{3}$$

where $P(x)\approx \delta p(x)$ (here δ represents a segment of the x domain which preferentially tends to 0). Equation (3) should be equal to S_Q^* in the limit of N going to infinity and $\delta\to 0$. Alternatively, the measure $P(x_i)$ relates to the distance δ (centered at x_i) which comprises a given number of nearest neighbors, n (originally $n=1$), or accordingly to the probability $\Pi_n(\delta)$ that the $(n-1)$ nearest neighbors have values x' within $x\pm\delta/2$ and the nth nearest neighbor is at a distance $\delta/2$ of x_i, i.e.,

*silvio.queiros@unilever.com; sdqueiro@gmail.com

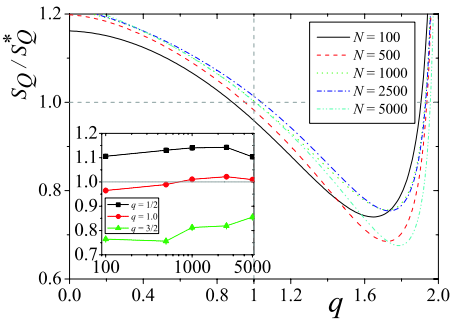

FIG. 1. (Color online) Ratio S_Q/S_Q^* vs the dual entropic parameter $q=2-Q$ for fixed $n=1$. The inset depicts the same ratio vs N for particular values of q. In this case the sets are composed of Gaussian distributed random variables.

$$\Pi_n(\delta) = \frac{(N-1)!}{(n-1)!\,(N-n-1)!}\frac{[P_i'(\delta)]^{n-1}}{[1-P_i'(\delta)]^{1+n-N}}\frac{dP_i'(\delta)}{d\delta}, \tag{4}$$

where $P_i'(\delta)=\int_{x-\delta/2}^{x+\delta/2} p_i'(z)dz$. Thence, we associate $\langle \ln_q P_i\rangle$ with $\overline{\ln_q P_i'}=\int \Pi_n(\delta)\ln_q P_i'(\delta)d\delta$ that yields [15]

$$\overline{\ln_q P_i'} = \frac{1-\dfrac{\Gamma[N]\Gamma[n+1-q]}{\Gamma[n]\Gamma[1+N-q]}}{q-1}. \tag{5}$$

Taking into consideration that $\ln_q(u\times v) = \ln_q u + \ln_q v + (1-q)\ln_q u \times \ln_q v$ and remembering that $P_i'(\delta)\approx p_i'\,\delta$ we obtain the final formula,

$$S_Q = \frac{\overline{\ln_q P_i'} - \langle \ln_q \delta\rangle}{1+(1-q)\langle \ln_q \delta\rangle}, \tag{6}$$

where $\langle \ln_q \delta\rangle$ represents the average of $\ln_q \delta$ over all points and samples accessible.

In practical terms, the algorithm is implemented the following way. For a fixed order of the vicinity, the distance $\delta/2$ from each point x_i of the data set under study to its nth nearest neighbor is determined. The values of δ are then used to compute the average of $\ln_q \delta$ that is used in the previous equation. The value of $\overline{\ln_q P_i'}$ is predefined when the values of q and n used in Eq. (5) are fixed.

Endowed with Eq. (6), we can rate the quality of the approximation by comparing its outcome with the predicted theoretical values given by Eq. (2). For the cases we will present hereinafter we have

$$S_{2-q}^* = \frac{1}{1-q} - \frac{(2\pi)^{(q-1)/2}}{(1-q)\sqrt{2-q}}, \tag{7}$$

for the Gaussian,

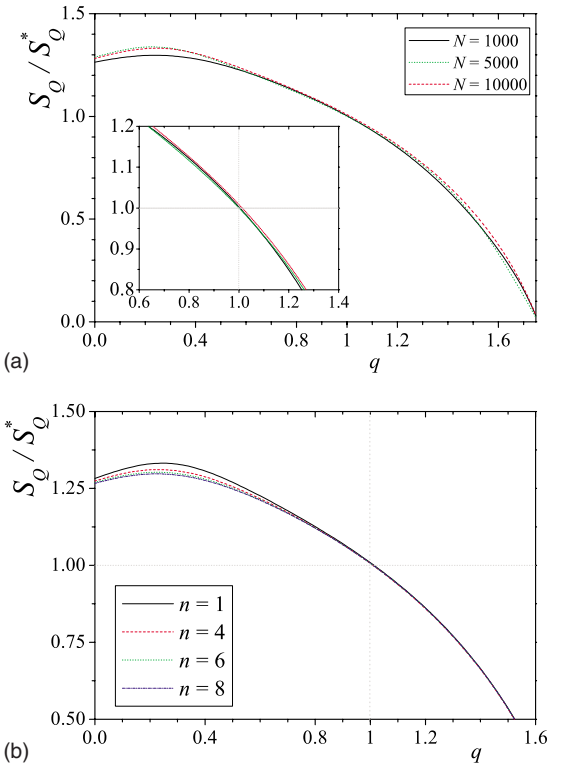

FIG. 2. (Color online) Upper panel: ratio S_Q/S_Q^* vs the dual entropic parameter $q=2-Q$ for fixed $n=1$. Lower panel: the same but for different n and fixed $N=5000$. In this case the sets are composed of Student-t (with 3 degrees of freedom) distributed random variables.

$$S_{2-q}^* = \frac{1}{1-q} - \frac{2^{2-q}3^{(q-1)/2}\Gamma\left[\dfrac{7}{2}-2q\right]}{\pi^{(3-q)/2}(1-q)\Gamma[4-2q]}, \tag{8}$$

for the Student-t with 3 degrees of freedom[1] and

$$S_{2-q}^* = -\ln_q 2, \tag{9}$$

for a uniform PDF defined between -1 and 1.

III. RESULTS

In order to test the actual efficiency of Eq. (6) we generated sets (typically 10^3) of random variables with a number of elements never larger than 10^4 on which we have applied the algorithm for diverse values of n.[2] The results depicted in Figs. 1–3 show that for the Gaussian and the Student-t, the Kozachenko-Leonenko approach is only a valuable estimator

[1]Because, under appropriate constraints, the entropy S_Q is maximized by the Student-t PDF, the latter has been also named Q-Gaussian distribution wherein the relation $Q=\frac{3+m}{1+m}$ between the entropic index, Q, and the degree of freedom m is valid [7].

[2]The random variables were bore by means of the extended cellular automata random number generator using the five-neighbor rule [16]. Additionally for the case of the Student-t we used the Bailey algorithm [17].

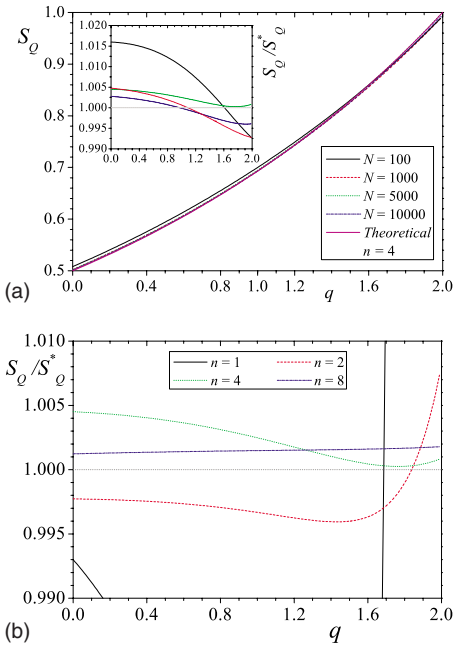

FIG. 3. (Color online) Upper panel: estimated S_Q vs the dual entropic parameter $q=2-Q$ (the inset represents the ratio S_Q/S_Q^* vs q). Lower panel: ratio S_Q/S_Q^* vs q for different n and $N=5000$. In this case the sets are composed of uniformly distributed random variables between -1 and 1 with the number of samples taken into account referred in the text.

for values of $q=1$, i.e., for the BG case, whereas for the uniform PDF it is quite effective.

For the Gaussian (see Fig. 1), we have verified that for $N<5000$ we have got error greater than 10% unless we are analyzing the $q=Q=1$ value. In this case, the error is already less than 1% for $N=100$. For the remaining $q\neq 1$ cases, we have not captured a monotonous behavior of the error and the ratio S_Q/S_Q^* with the number of elements of the set or the order of the neighbor used. In respect of the dependence of S_Q/S_Q^* on n (for fixed N), we have verified alike behavior with $n=1$ which presents the best estimations for any fixed N tested.

Regarding the Student-t case, we have noticed the same qualitative results; i.e., the KLA algorithm tends to overestimate (underestimate) the entropy S_Q^* for $Q>1$ ($Q<1$) independently of the size of the series and the order of the nearest neighbor taken into reference.[3] Once again, for the case $Q=1$ the algorithm caters for an excellent approach even for relatively small sets ($N<1000$) as exhibited in Fig. 2.

As shown in Fig. 3, the ineffectiveness we have reported so far is only challenged when the uniform PDF is considered. In this case, for values of $n>1$, we have verified that the KLA is a trustworthy estimator of the theoretical entropy of a system. For instance, by considering sets of 100 variables we have achieved discrepancies never greater than 2%. Comparing the KLA results with entropy evaluations obtained by a simple binning of the sets we verify the algorithm

[3]Although only $n=1$ is shown herein, we let n run up to the remote value of $n=100$.

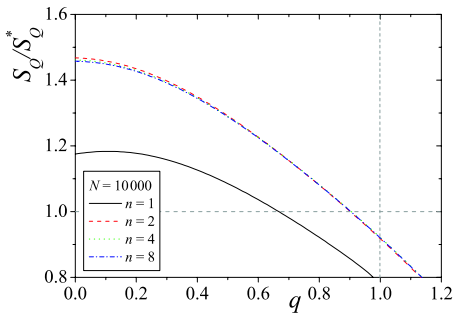

FIG. 4. (Color online) Ratio S_Q/S_Q^* vs the dual entropic parameter $q=2-Q$ for fixed $N=10\,000$. In this case the sets are composed of stochastic Feller-like process as described in the text.

is only slightly better than the latter approach. Taking into account the computation time we would say that the KLA does not pay off.

Complementary, we now study the effectuation of the KLA to time series generated in two different ways [18]. First, we consider the stochastic differential equation $dx=-\gamma x dt+\sqrt{\theta[P(x)]^{\nu}}dW_t$ (Itô notation) [19] whose stationary PDF is the q-Gaussian. Additionally, the process can reproduce at the first level the intraday dynamics of the price fluctuations of some financial markets. We have used $\gamma=100^{-1}$, $\theta=\gamma\sqrt{2/\pi}$, and $\nu=-1/2$ which yields the $m=3$ Student-t [$(q=1.5)$-Gaussian] as the stationary PDF. This case is marked by the existence of linear correlations between the variables which affect the quality of the estimation as plotted in Fig. 4. Despite the fact that the best estimative is still for values of q close to 1, the KLA is not so accurate as in the independent case. Nevertheless, we can surmount this situation taking into consideration that a shuffling procedure does not alter the stationary PDF of stationary process.

The second case corresponds to time series generated by a heteroskedastic process enclosed within the fractional autoregressive conditional heteroskedasticity class in which discrete stochastic variables $x_t=\sigma_t\omega_t$ (ω_t follows a Gaussian) are generated with $\sigma_t^2=a+b\sum_{i=t_0}^{t-1}\mathcal{K}(i-t+1)x_i^2$, where $\mathcal{K}(t')\sim\exp_\zeta[t']$ ($t'\leq 0$, $T>0$) [20] and $\exp_\zeta(\cdots)$ is the inverse function of $\ln_\zeta(\cdots)$. In spite of generating uncorrelated variables, this model exhibits long-lasting correlations in the variance (nonlinear dependence for x) and its probabilistic analysis provides strong statistical evidence that the stationary PDF is a Student-t. Using $\zeta=1.375$, $b=0.9375$, and $a=1-b$ we have obtained a $(q=1.54)$-Gaussian. Employing the KLA algorithm, we have obtained equivalent results to the previous linearly correlated case (see Fig. 5). We must be careful and mind the fact that the resulting PDF is not exact though. It should be noted that the error in the entropy estimation is greater than the error presented in the adjustment by a q-Gaussian.

IV. REMARKS

In this Brief Report we have introduced a generalization of the well-known binless Lozachenko-Leonenko entropy estimator to appraise the (Tsallis) nonadditive entropy in sys-

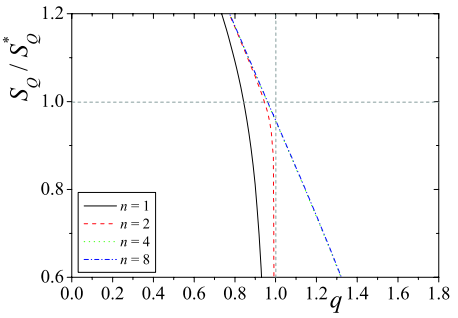

FIG. 5. (Color online) Estimated S_Q vs the dual entropic parameter $q=2-Q$ for fixed $N=10\,000$. In this case the sets are composed of ($q=1.54$)-Gaussian (approximately) generated according to the heteroskedastic process described in the text.

tems with a small number of observations for which binning strategies are likely to present strong deviation from the expected theoretical result. By comparing numerical results with theoretical values we have verified that the KLA approach is not effective. Although we do not have any irrefutable reasoning which explains the results reported herein above, we believe that they are a demonstration of the bias introduced by the Q entropic index in the weight the probability $p(x)$ in Eq. (1) [21]. Explicitly, for values of $Q>1$ ($q<1$) we have $[p(x)]^Q>p(x)$ if $p(x)>1$ and $[p(x)]^Q<p(x)$ otherwise. On the other hand, if $Q<1$ ($q>1$) we have $[p(x)]^Q<p(x)$ if $p(x)>1$ and $[p(x)]^Q>p(x)$ if $p(x)<1$. Apparently, this bias is overestimated for $q<1$ and underestimated for $q>1$ by the evaluation of the $\langle \ln_q \delta \rangle$. In the case of uniform PDF, the bias is shed and the KLA yields a remarkable result. For $q=1$, the accuracy of the algorithm only diminishes when dependent time series are analyzed.

Regarding the Renyi entropic form we have mentioned, $S_R=\{\ln \int [p(x)]^\alpha dx\}/(1-\alpha)$ ($\alpha \geq 0$), a similar approach can be implemented, albeit a description involving averages similar to Eqs. (2) and (3) is nontrivial. Nonetheless, allowing for the fact that at the first order $S_R=S_Q$ ($\alpha=Q$), further work should deem whether the remaining terms in the expansion of S_R either set off the error presented by the first approximation (leading to the effectiveness of the KLA) or sum up to it. Overall, bearing in mind its importance for a reliable study of many complex phenomena, it is expected that new binless or binning strategies [10] for the evaluation of entropic functionals such as S_Q will correct the shortcoming conveyed here by the KLA approach.

ACKNOWLEDGMENT

This work benefited from support of the Marie Curie Programme (European Union).

[1] E. G. D. Cohen, in *Boltzmann and Statistical Mechanics*, Proceedings of the International Meeting Boltzmann's Legacy 150 Years After His Birth (Atti della Accademia Nazionale dei Lincei, Rome, 1997), pp. 9–23.
[2] K. Huang, *Statistical Mechanics* (Wiley, New York, 1963).
[3] K. L. Volkovyski and Ya. G. Sinai, Funct. Anal. Appl. **5**, 185 (1971).
[4] C. Tsallis, M. Gell-Mann, and Y. Sato, Proc. Natl. Acad. Sci. U.S.A. **102**, 15377 (2005).
[5] A. Renyi, *Probability Theory* (North-Holland, Amsterdam, 1970).
[6] C. Tsallis, J. Stat. Phys. **52**, 479 (1988).
[7] A. M. C. de Souza and C. Tsallis, Physica A **236**, 52 (1997).
[8] C. Tsallis, *Introduction to Nonextensive Statistical Mechanics: Approaching a Complex World* (Springer, Berlin, 2009); a comprehensive list of applications is also available at http://tsallis.cat.cbpf.br/biblio.htm
[9] R. Abramov, J. Comput. Phys. **226**, 621 (2007); **228**, 96 (2009); R. Abramov and A. J. Majda, SIAM J. Sci. Comput. (USA) **26**, 411 (2004).
[10] R. Q. Quiroga, T. Kreuz, and P. Grassberger, Phys. Rev. E **66**, 041904 (2002).
[11] A. Kraskov, H. Stogbauer, and P. Grassberger, Phys. Rev. E **69**, 066138 (2004).
[12] F. Caruso and C. Tsallis, Phys. Rev. E **78**, 021102 (2008).
[13] S. M. Duarte Queirós, Physica D **238**, 764 (2009).
[14] L. F. Kozachenko and N. N. Leonenko, Probl. Inf. Transm. **23**, 95 (1987); P. Grassberger, Phys. Lett. A **107**, 101 (1985); J. D. Victor, Phys. Rev. E **66**, 051903 (2002).
[15] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series and Products* (Academic Press, New York, 1980).
[16] J. E. Gentle, *Random Number Generation and Monte Carlo Methods*, 2nd ed. (Wiley, New York, 1995), Vol. 2.
[17] R. W. Bailey, Math. Comput. **62**, 779 (1994).
[18] See EPAPS Document No. E-PLEEE8-80-126911 for supplementary information on the stochastic models used. For more information on EPAPS, see http://www.aip.org/pubservs/epaps.html.
[19] L. Borland, Phys. Rev. E **57**, 6634 (1998).
[20] S. M. Duarte Queirós, EPL **80**, 30005 (2007).
[21] C. Tsallis, Braz. J. Phys. **29**, 1 (1999).