Generalized entropy-based criterion for consistent testing

Constantino Tsallis
Centro Brasileiro de Pesquisas Físicas, Rua Xavier Sigaud 150, 22290-180 Rio de Janeiro, Rio de Janeiro, Brazil

(Received 19 March 1998)

Through the use of a recently introduced, nonextensive, entropy, we generalize that of Kullback and Leibler [Ann. Math. Stat. **22**, 79 (1961)] and study its properties. This in turn enables the proposal of a consistent criterion for testing relevant hypotheses such as the independence of random variables. Straightforward applications are shown to be possible for (physical, geophysical, economic, and biological) time series. [S1063-651X(98)03808-2]

PACS number(s): 05.20.−y, 05.40.+j, 02.50.Wp

The problem of consistent testing, i.e., *discrimination between two hypotheses*, is a central one in as varied areas as physics (e.g., in high-energy elementary particles experiments), geophysics (e.g., rainfall times series and El Niño climatological phenomena), economics (e.g., degree of correlation in time series of quantities of financial interest), and biology (e.g., correlations in nucleotides of DNA chains and cardiological and electroencephalographic rhythms), just to mention a few. Nonparametric testing is of course a very well justified one and, on an entropy basis, has been proposed and used by several authors [1–3]. Several years ago, Robinson [4] used the Kullback-Leibler measure of information [5] [which is based on the Boltzmann-Gibbs-Shannon (BGS) entropy] to make an elegant discussion of *independence versus dependence* in time series of (daily, weekly, and monthly) exchange rates of several important currencies against the U.S. dollar. More precisely, he used data of the Bank of England covering the period 2 January 1978 through 28 June 1985. It is probably unnecessary to say that physical, geophysical, biological, and other time series could usefully be processed in the same manner.

On a quite different vein, we proposed several years ago [6] a generalization of the usual Boltzmann-Gibbs statistical mechanics, hence of thermodynamics itself. This generalization addresses *nonextensive* systems (long-range interactions, long-range microscopic memory, fractal or multifractal relevant space-time, etc.) and is based on the entropic form [written here for a continuous random variable characterized by the probability distribution $p(x)$]:

$$S_q(p) \equiv -\int dx\, p(x)\,\frac{[p(x)]^{q-1}-1}{q-1} = -\int dx[p(x)]^q \frac{[p(x)]^{1-q}-1}{1-q} \times\left(\int dx\, p(x)=1; q\in\mathbb{R}\right), \tag{1}$$

which (using $[p(x)]^{q-1}\sim 1+(q-1)\ln p(x)$) recovers the usual BGS entropy $S_1(p)\equiv -\int dx\, p(x)\ln p(x)$ in the limit $q\rightarrow 1$. The entropic index q characterizes the degree of nonextensivity reflected in the (easily verified) pseudoadditivity property $S_q(A+B)=S_q(A)+S_q(B)+(1-q)S_q(A)S_q(B)$, where A and B are two *independent* systems in the sense that the probability distribution of $A+B$ *factorizes* into those of A and of B.

This generalization retains much of the formal structure of the standard theory such as the Legendre thermodynamic structure, H theorem, Onsager reciprocity theorem, Kramers and Wannier relations, Bogolyubov inequality, and thermodynamic stability, among others [6,7], and has been applied to many anomalous physical systems. Within a long list we may mention Lévy and correlated anomalous diffusions (see [8] and references therein), stellar polytropes [9,10], pure-electron plasma two-dimensional turbulence [10], solar neutrinos [11], anomalous phonon-electron thermalization in ion-bombarded solids [12], peculiar velocities of galaxies [13], inverse bremsstrahlung in plasma [14], cosmology [15], nonlinear dynamical low-dimensional (at the edge of chaos) [16] as well as high-dimensional (at self-organized criticality [17]) [18] dissipative systems, long-range-interacting fluids, and magnets [19].

The aim of the present work is to show how these ideas can be used to propose, along Robinson's lines, a generalized consistent testing, which could be useful for handling a great variety of problems. Let us first recall the Kullback-Leibler measure of information (or *cross entropy* or *relative entropy* or *mutual information*)

$$I_1(p,p_0)\equiv\int dx\, p(x)\ln\frac{p(x)}{p_0(x)} = -\int dx\, p(x)\ln\frac{p_0(x)}{p(x)}, \tag{2}$$

where $p_0(x)$ is the so-called *reference* (or *default*) *distribution* (uniform, Gaussian, Lorentzian, and Poisson distributions are common choices) and the meaning of the subindex 1 will become transparent in a little while. By using that $\ln r \geqslant 1-(1/r)$ $[r\equiv p(x)/p_0(x)>0]$, it is easily seen that this quantity satisfies

$$I_1(p,p_0)\geqslant 0 \quad \forall(p,p_0). \tag{3}$$

$I_1(p,p_0)=0$ if and only if $p=p_0$ almost everywhere. Property (3) must be emphasized since it constitutes the very basis for *consistency* of the present nonparametric testing. Indeed, $I_1(p,p_0)$ can be used as a *distance* of p with regard to p_0 [notice that, unless $p=p_0$, generically $I_1(p,p_0)\neq I_1(p_0,p)$, a property to which we shall return later on].

Another important property of $I_1(p,p_0)$ is that it is form invariant under variable transformation. Indeed, if we perform the variable transformation $x=f(y)$, the measure preservation implies that $p(x)dx=\tilde{p}(y)dy$, where $\tilde{p}(y)$ is the new distribution law [and analogously for $p_0(x)$]. Since $p/p_0=\tilde{p}/\tilde{p}_0$, $I_1(p,p_0)=\int dy\, \tilde{p}(x(y))\ln[\tilde{p}(x(y))/\tilde{p}_0(x(y))] =I_1(\tilde{p},\tilde{p}_0)$, which proves the above-mentioned form invariance. As a last important property let us mention that, if we choose as $p_0(x)$ a *uniform* distribution on a compact support of length W, then it is straightforward to verify that

$$I_1(p,1/W)=\ln W-S_1(p), \tag{4}$$

which presents the Kullback-Leibler entropy as the departure of the BGS entropy from its value at equiprobability.

The definition of $I_1(p,p_0)$ and the generalized entropic form $S_q(p)$ [Eq. (1)] naturally lead to the generalization

$$\begin{aligned} I_q(p,p_0) &\equiv \int dx\, p(x)\frac{[p(x)/p_0(x)]^{q-1}-1}{q-1} \\ &= -\int dx\, p(x)\frac{[p_0(x)/p(x)]^{1-q}-1}{1-q}, \end{aligned} \tag{5}$$

where we can immediately verify that the limit $q\rightarrow 1$ recovers the standard Kullback-Leibler entropy (2). Let us now generalize (by following along the lines of [22]) the very important property (3). With $r>0$, we have that

$$\begin{aligned} \frac{r^{q-1}-1}{q-1} &\geqslant 1-\frac{1}{r} \quad \text{if} \quad q>0 \\ &= 1-\frac{1}{r} \quad \text{if} \quad q=0 \\ &\leqslant 1-\frac{1}{r} \quad \text{if} \quad q<0 \end{aligned} \tag{6}$$

(for $q\neq 0$, the equality holds if and only if $r=1$). Consequently, for, say, $q>0$, we have that

$$\frac{\left[\dfrac{p(x)}{p_0(x)}\right]^{q-1}-1}{q-1} \geqslant 1-\frac{p_0(x)}{p(x)}; \tag{7}$$

hence

$$\begin{aligned} \int dx\, p(x)\frac{\left[\dfrac{p(x)}{p_0(x)}\right]^{q-1}-1}{q-1} &\geqslant \int dx\, p(x)\left[1-\frac{p_0(x)}{p(x)}\right] \\ &= 1-1=0. \end{aligned} \tag{8}$$

However, the left-hand side member of this inequality is precisely $I_q(p,p_0)$. Consequently, Eqs. (6) imply

$$\begin{aligned} I_q(p,p_0) &\geqslant 0 \quad \text{if} \quad q>0 \\ &= 0 \quad \text{if} \quad q=0 \\ &\leqslant 0 \quad \text{if} \quad q<0. \end{aligned} \tag{9}$$

For $q\neq 0$, the equalities hold if and only if $p=p_0$ almost everywhere. Equation (3), as well as the above-mentioned form invariance, is thus generalized for arbitrary q. By performing the transformation $q-\frac{1}{2}\leftrightarrow\frac{1}{2}-q$ in the definition (5) we can prove that

$$\frac{I_q(p,p_0)}{q}=\frac{I_{1-q}(p_0,p)}{1-q}. \tag{10}$$

Consequently, as a family of entropy-based testings, it is enough to consider $q\geqslant\frac{1}{2}$, for which

$$I_q(p,p_0)\geqslant 0, \tag{11}$$

the equality holding if and only if $p=p_0$ almost everywhere. The criterion indicated in Eq. (9) implies, for the particular case $q=\frac{1}{2}$,

$$\int dx\sqrt{p(x)p_0(x)}\leqslant 1. \tag{12}$$

This expression can be interpreted as the continuous version of the scalar product between two unitary vectors, namely, $\sqrt{p(x)}$ and $\sqrt{p_0(x)}$, and is directly related to the so-called *Fisher genetic distance* [20].

For the particular case $q=2$, the criterion (9) becomes

$$\int dx[p(x)]^2/p_0(x)\leqslant 1. \tag{13}$$

Also, except for $I_{1/2}$ (and the trivial case I_0), we easily see that $I_q(p_0,p)\neq I_q(p,p_0)$ unless $p=p_0$ almost everywhere. Consequently, if for some reason we want a *reciprocal* ''distance'' between p and p_0, it might be convenient to define a symmetrized quantity such as

$$I_q^S(p,p_0)\equiv\frac{1}{2}[I_q(p,p_0)+I_q(p_0,p)]; \tag{14}$$

hence $I_q^S(p,p_0)=I_q^S(p_0,p) \quad \forall(p,p_0,q)$.

As a last property, let us generalize Eq. (4). By choosing, once again, as $p_0(x)$ the uniform distribution on a compact support of length W, we easily establish that

$$I_q(p,1/W)=\frac{W^{1-q}-1}{1-q}-W^{q-1}S_q(p). \tag{15}$$

Let us now adapt the main results of this paper to the problem of independence of random variables. Let us consider the two-dimensional random variable $z\equiv(x,y)$ and its corresponding distribution function $p(x,y)$ with $\int dx\,dy\,p(x,y)=1$. The marginal distribution functions are then given by $h_1(x)\equiv\int dy\,p(x,y)$ and $h_2(y)\equiv\int dx\,p(x,y)$. In this situation, the discrimination criterion for independence of course concerns the comparison of $p(x,y)$ with $p_0(x,y)\equiv h_1(x)h_2(y)$. The one-dimensional random variables x and y are independent if and only if $p(x,y) =p_0(x,y)[\forall(x,y)]$. The criterion (11) becomes

$$\int dx\,dy\,p(x,y)\frac{\left[\frac{p(x,y)}{h_1(x)h_2(y)}\right]^{q-1}-1}{q-1}\geqslant 0\left(q\geqslant\frac{1}{2}\right). \tag{16}$$

The evaluation of this quantity gives a satisfactory measure of the degree of dependence between x and y; when and only when it vanishes, x and y can be considered independent. In the $q\to 1$ limit, this criterion becomes the usual one (see, for instance, [4])

$$\int dx\,dy\,p(x,y)\ln p(x,y)-\int dx\,h_1(x)\ln h_1(x)$$
$$-\int dy\,h_2(y)\ln h_1(y)\geqslant 0. \tag{17}$$

For $q=1/2$ we have

$$\int dx\,dy\sqrt{p(x,y)h_1(x)h_2(y)}\leqslant 1. \tag{18}$$

The particular case $q=2$ becomes

$$\int dx\,dy\frac{[p(x,y)]^2}{h_1(x)h_2(y)}\geqslant 1. \tag{19}$$

This can be considered as a satisfactory ''quadratic'' criterion, as opposed to the quantity basically introduced in [21] (for the particular case $h_1=h_2\equiv h$),

$$\int dx\,dy[p(x,y)]^2-\left(\int dx[h(x)]^2\right)^2. \tag{20}$$

Indeed (see also [4]), this quantity has no definite sign and its zero value does not guarantee an independence between x and y. In other words, it cannot be considered as an optimal criterion and could, in principle, very well be replaced by the present criterion (19).

If for a particular use we have reasons to prefer a symmetrized criterion, we can replace Eq. (16) by

$$I_q^S(p(x,y),h_1(x)h_2(y))\geqslant 0 \quad (q\geqslant\tfrac{1}{2}). \tag{21}$$

The generalization for an arbitrary number d of variables (with $d\geqslant 2$) is straightforward:

$$I_q^S(p(x_1,x_2,\ldots,x_d),p_0(x_1,x_2,\ldots,x_d))\geqslant 0 \quad (q\geqslant\tfrac{1}{2}), \tag{22}$$

with

$$p_0(x_1,x_2,\ldots,x_d)$$
$$\equiv\left[\int dx_2dx_3\cdots dx_d p(x_1,x_2,\ldots,x_d)\right]$$
$$\times\left[\int dx_1dx_3\cdots dx_d p(x_1,x_2,\ldots,x_d)\right]$$
$$\times\cdots\times\left[\int dx_1dx_2\cdots dx_{d-1}p(x_1,x_2,\ldots,x_d)\right]. \tag{23}$$

The equality in Eq. (22) holds if and only if $(x_1,x_2,\ldots,x_d)$ can all be considered independent.

Let us finally make the bridge with a (physical, geophysical, economical, and biological) time series denoted $\{\xi_t\}$ with $t=0,1,2,\ldots$. One can, for instance, define [4] $X_t\equiv\ln(\xi_t/\xi_{t-1})$ and use $z\equiv(x,y)\equiv(X_t,X_{t-1})$, i.e., a $d=2$ problem. It is obvious that, according to the specific problem, it might be useful to work on larger spaces (i.e., $d>2$).

Summarizing, by following along the lines of the recently formulated nonextensive entropy and thermostatistics [6], we have established, on firm mathematical grounds, a *generalized criterion for consistent testing of independence between random variables*, which we propose as a practical tool for analyzing data such as DNA or peptide sequences and all types of computational or experimental time series. The results depend upon the entropic index q: It is expected that, for every specific use, better discrimination will be achieved with appropriate ranges of values of q. This was indeed the case of a recent wavelet-entropy analysis [23] of electroencephalographic data of epileptic turtles and human patients; the best values for clinical analysis turned out to be in the neighborhood of $q=5$. The value of q in the vicinity of which the criterion will be more fruitful no doubt is related to the (multi)fractal structure of the signal(s) under study, which in turn reflects the deep microscopic or mesoscopic (generically nonlinear) dynamics in the phase space of the system. The ubiquitous, so-called *complex systems* possibly are ideal candidates for a variety of applications. At the present moment, the analysis along these lines of the El Niño data is in progress. Several interesting effects emerge as a function of q that will be presented elsewhere.

I acknowledge R. Guenzburger for calling my attention to Robinson's work [4] and A. R. Plastino and L. Borland for fruitful discussions. This work was partially supported by PRONEX (Brazil).

[1] Y. G. Dmitriev and F. P. Tarasenko, Theor. Probab. Appl. **18**, 628 (1973).

[2] I. A. Ahmad and P.-I. Lin, IEEE Trans. Inf. Theory **22**, 372 (1976).

[3] O. Vasicek, J. R. Stat. Soc., Ser. B **38**, 54 (1976).

[4] P. M. Robinson, Rev. Econ. Studies **58**, 437 (1991).

[5] S. Kullback and R. A. Leibler, Ann. Math. Stat. **22**, 79 (1961); S. L. Braunstein, Phys. Lett. A **219**, 169 (1996), and references therein.

[6] C. Tsallis, J. Stat. Phys. **52**, 479 (1988); E. M. F. Curado and C. Tsallis, J. Phys. A **24**, L69 (1991); **24**, 3187(E) (1991); **25**, 1019 (1992); C. Tsallis, Phys. Lett. A **206**, 389 (1995); see http://tsallis.cat.cbpf.br/biblio.htm for an updated bibliography on the subject.

[7] A. M. Mariz, Phys. Lett. A **165**, 409 (1992); M. O. Caceres, *ibid.* **218**, 471 (1995); A. K. Rajagopal, Phys. Rev. Lett. **76**, 3469 (1996); A. Chame and E. V. L. de Mello, Phys. Lett. A **228**, 159 (1997); E. K. Lenzi, L. C. Malacarne, and R. S.

Mendes, Phys. Rev. Lett. **80**, 218 (1998); A. K. Rajagopal, R. S. Mendes, and E. K. Lenzi, Phys. Rev. Lett. **80**, 3907 (1998).

[8] M. F. Shlesinger, G. M. Zaslavsky, and U. Frisch, *Levy Flights and Related Topics in Physics* (Springer, Berlin, 1995); D. H. Zanette and P. A. Alemany, Phys. Rev. Lett. **75**, 366 (1995); **77**, 2590 (1996); M. O. Caceres and C. E. Budde, *ibid.* **77**, 2589 (1996); C. Tsallis, S. V. F. Levy, A. M. C. de Souza, and R. Maynard, *ibid.* **77**, 5422 (1996); **77**, 5442(E) (1996); A. R. Plastino and A. Plastino, Physica A **222**, 347 (1995); C. Tsallis and D. J. Bukman, Phys. Rev. E **54**, R2197 (1996).

[9] A. R. Plastino and A. Plastino, Phys. Lett. A **174**, 384 (1993).

[10] B. M. Boghosian, Phys. Rev. E **53**, 4754 (1996).

[11] G. Kaniadakis, A. Lavagno, and P. Quarati, Phys. Lett. B **369**, 308 (1996); P. Quarati, A. Carbone, G. Gervino, G. Kaniadakis, A. Lavagno, and E. Miraldi, Nucl. Phys. A **621**, 345c (1996).

[12] I. Koponen, Phys. Rev. E **55**, 7759 (1997).

[13] A. Lavagno, G. Kaniadakis, M. Rego-Monteiro, P. Quarati, and C. Tsallis, Astrophys. Lett. and Commun. **35**, 449 (1998).

[14] C. Tsallis and A. M. C. de Souza, Phys. Lett. A **235**, 444 (1997).

[15] V. H. Hamity and D. E. Barraco, Phys. Rev. Lett. **76**, 4664 (1996); D. F. Torres, H. Vucetich, and A. Plastino, *ibid.* **79**, 1588 (1997).

[16] C. Tsallis, A. R. Plastino, and W.-M. Zheng, Chaos Solitons Fractals **8**, 885 (1997); U. M. S. Costa, M. L. Lyra, A. R. Plastino, and C. Tsallis, Phys. Rev. E **56**, 245 (1997); M. L. Lyra and C. Tsallis, Phys. Rev. Lett. **80**, 53 (1998).

[17] P. Bak, C. Tang, and K. Wiesenfeld, Phys. Rev. Lett. **59**, 381 (1987).

[18] F. A. Tamarit, S. A. Cannas, and C. Tsallis, Eur. J. Phys. B **1**, 545 (1998); A. R. R. Papa and C. Tsallis, Phys. Rev. E **57**, 3923 (1998).

[19] P. Jund, S. G. Kim, and C. Tsallis, Phys. Rev. B **52**, 50 (1995); C. Tsallis, Fractals **3**, 541 (1995); J. R. Grigera, Phys. Lett. A **217**, 47 (1996); S. A. Cannas and F. A. Tamarit, Phys. Rev. B **54**, R12 661 (1996); S. A. Cannas and A. C. N. Magalhaes, J. Phys. A **30**, 3345 (1997); L. C. Sampaio, M. P. de Albuquerque, and F. S. de Menezes, Phys. Rev. B **55**, 5611 (1997); S. E. Curilef, Ph.D. thesis, Centro Brasileiro de Pesquisas Fı́sicas, 1997 (unpublished); C. Anteneodo and C. Tsallis, Phys. Rev. Lett. **80**, 5313 (1998).

[20] W. K. Wootters, Phys. Rev. D **23**, 357 (1981).

[21] W. Brock, D. Dechert, and J. Scheinkman (unpublished).

[22] A. Plastino and C. Tsallis, J. Phys. A **26**, L839 (1993).

[23] L. G. Gamero, A. Plastino, and M. E. Torres, Physica A **246**, 487 (1997); A. Capurro, L. Diambra, D. Lorenzo, O. Macadar, M. T. Martin, A. Plastino, E. Rofman, M. E. Torres, and J. Velluti, INRIA Report No. 3184, 1997 (unpublished).