Physica A 510 (2018) 68–76

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Generalized divergences from generalized entropies

Leonardo E. Riveaud [a], Diego Mateos [b], Steeve Zozor [c], Pedro W. Lamberti [d,c,*]

[a] *CONICET, Rivadavia 1917, C1033AAJ, CABA, Argentina*
[b] *Facultad de Ciencias y Técnica, Universidad Autónoma de Entre Ríos, Entre Ríos, Avenida Ramirez 1143, ZC3100, Argentina*
[c] *Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France*
[d] *CONICET and Facultad de Matemática, Astronomia, Fisica y Computación, Universidad Nacional de Córdoba, Ciudad Universitaria, X5000HUA, Córdoba, Argentina*

HIGHLIGHTS

- An alternative road to Bregman's divergences.
- Divergences from (h, ϕ) entropies.
- Weighted Bregman's divergences and some applications to the study of time series.

ARTICLE INFO

Article history:
Received 8 May 2018
Received in revised form 18 June 2018
Available online xxxx

Keywords:
Divergences
Information theory
Time series

ABSTRACT

Several quantifiers of information, also known as entropies, have been introduced in different contexts and from different motivations. For almost each one of these entropies, a measure of the loss (or gain) of information has been introduced. In this work we introduce generalized weighted divergences associated with an arbitrary entropy. The resulting measures are closely related to Bregman divergences. We study the main formal properties of the resulting divergences, we extend them to weighted probability distributions and we apply some of them to the analysis of simulated and real time series.

1. Introduction

Searching for distances and similarity (or dissimilarity) measures between probability distributions, also called divergences, is a topic of great interest in pure and applied mathematical statistics. It is well known that not all distances or divergences are adequate for the treatment of every problem. Therefore, having a variety of divergences could be useful both for theoretical studies as well as in the context of applications. One widely used similarity measure is the Kullback–Leibler divergence (KLD) which has the expression

$$K(\mathcal{P}, \mathcal{Q}) = \sum_i p_i \log_2 \left(\frac{p_i}{q_i}\right) \tag{1}$$

where $\mathcal{P} = \{p_i\}_{i=1}^N$ and $\mathcal{Q} = \{q_i\}_{i=1}^N$ represent probability distributions for a N-states discrete random variable X. This divergence is naturally related to the Shannon entropy, $H^S(\mathcal{P}) = -\sum_i p_i \log_2 p_i$ [1]. The KLD has several interesting interpretations in the realm of Information Theory (IT). For example, if one is concerned about the length of a code to

[*] Corresponding author at: CONICET and Facultad de Matemática, Astronomia, Fisica y Computación, Universidad Nacional de Córdoba, Ciudad Universitaria, X5000HUA, Córdoba, Argentina.
E-mail address: pwlamberti@yahoo.com (P.W. Lamberti).

https://doi.org/10.1016/j.physa.2018.06.111
0378-4371/© 2018 Published by Elsevier B.V.

represent the random variable X, and one uses the distribution $\mathcal{Q}$ to make the code instead of the true distribution $\mathcal{P}$, the average description length for the code is given by the amount $H^S(\mathcal{P}) + K(\mathcal{P}, \mathcal{Q})$ instead of $H^S(\mathcal{P})$ [2].

The KLD is definite positive, non symmetric and not bounded. It belongs to a broad class of dissimilarity measures, known as Csiszár divergences [3]. These measures have the form:

$$D_f(\mathcal{P}, \mathcal{Q}) = \sum_i p_i f\left(\frac{q_i}{p_i}\right) \tag{2}$$

with f a convex function. Additionally, we concentrate on the case where f is of class C^2 and such that $f(1) = 0$ and $f''(1) = 1$.[1] For the KLD, the corresponding function is $f(x) = -\log_2 x$. Csiszár divergences have been applied to the study of several physical phenomena, particularly in the context of non-equilibrium thermo-statistics [4].

Every Csiszár divergence with f of class C^2 has a second order Taylor expansion of the form

$$D_f\left(\mathcal{P}, \mathcal{P} + \delta\mathcal{P}\right) = \frac{1}{2}\sum_i \frac{\delta p_i^2}{p_i} + o\left(\sum_i \delta p_i^2\right) \tag{3}$$

where $\mathcal{P}$ and $\mathcal{P} + \delta\mathcal{P}$ are two "close" distributions and o is a negligible reminder.

Since the emergence of IT several entropies have been studied. Two notorious cases are the Rényi entropy [5] and the Havrda–Charvát entropy (HC) [6]. The first one is given by

$$H_\alpha^R(\mathcal{P}) = \frac{1}{1-\alpha}\log_2\left(\sum_i p_i^\alpha\right) \tag{4}$$

with α a real parameter. The Havrda–Charvát entropy, also known in physics as the Tsallis entropy has the expression

$$H_q^{HC}(\mathcal{P}) = \frac{k}{q-1}\left(1 - \sum_i p_i^q\right) \tag{5}$$

with q a real parameter and k a proportionality factor. In both cases they encompass the Shannon entropy as a special case, in the limit when $\alpha \to 1$ and $q \to 1$ respectively. Others authors have introduced families of entropies that have the Rényi entropy and the Havrda–Charvát one as particular cases. A pioneering work in that direction was published by Daroczy [7] and later extended by Burbea and Rao [8] or Salicrú et al. [9] for instance.

Usually some basic properties are required for an entropy—that we will denote generically as H^G:

[E1] to be continuous with respect to (w.r.t). the p_i's;
[E2] to be non-negative;
[E3] to be equal to zero in the deterministic case, i.e., when one of the $p_i = 1$ and the others are equal to zero;
[E4] to reach the maximum value for the uniform distribution, i.e., when $p_i = \frac{1}{N}\ \forall\, i$, and
[E5] to be a concave function w.r.t. $\mathcal{P}$ in the sense of satisfying

$$H^G\left(\sum_k \lambda_k \mathcal{P}_k\right) \ge \sum_k \lambda_k H^G\left(\mathcal{P}_k\right), \qquad \forall\, \lambda_k > 0 \quad \text{such that} \quad \sum_k \lambda_k = 1 \quad \text{and} \quad \forall\, \mathcal{P}_k$$

This last condition is satisfied for the Shannon entropy, for the HC entropies when $q > 0$, and, in the Rényi case when $0 < \alpha < 1$.

Let us observe that every entropy satisfying condition [E5] allows to define a (Jensen-like)-divergence measure through the expression:

$$D_J^{H^G}\left(\mathcal{P}, \mathcal{Q}; \frac{1}{2}, \frac{1}{2}\right) \equiv H^G\left(\frac{1}{2}\mathcal{P} + \frac{1}{2}\mathcal{Q}\right) - \frac{1}{2}H^G\left(\mathcal{P}\right) - \frac{1}{2}H^G\left(\mathcal{Q}\right) \tag{6}$$

This is a definite positive quantity, equal to zero if and only if $\mathcal{P} = \mathcal{Q}$, bounded, and symmetric. Even more, it can be easily extended to an arbitrary number of distributions $\{\mathcal{P}_k\}_{k=1}^K$ which could have assigned different weights $\{\pi_k \ge 0\}_{k=1}^K$ such that $\sum_k \pi_k = 1$:

$$D_J^{H^G}\left(\mathcal{P}_1, \ldots, \mathcal{P}_K\,;\ \pi_1, \ldots, \pi_K\right) \equiv H^G\left(\sum_{k=1}^K \pi_k \mathcal{P}_k\right) - \sum_{k=1}^K \pi_k H^G\left(\mathcal{P}_k\right) \tag{7}$$

These divergences were introduced by Burbea and Rao in [8]. For the case of Shannon entropy the corresponding measure (7) it is known as the Jensen–Shannon divergence (JSD). Among the properties of the JSD, stands out that it can be interpreted in the context of Bayesian inference [10].

[1] This is not a restriction, except that f'' must be non-zero in 1. Indeed, if f is convex, $\frac{f(x)-x}{f''(1)}$ remain convex ($f''(1) > 0$ from the convexity of f), so that the divergence is defined up to a shift and scaling factor.

For the particular case of two probability distributions the JSD has a metric character. In fact, it has been possible to prove that the power $\left[D_J^S\left(\mathcal{P},\mathcal{Q};\frac{1}{2},\frac{1}{2}\right)\right]^r$ with $r\in\left(0,\frac{1}{2}\right]$ satisfies the triangle inequality [11]. Therefore, it provides of a mono parametric family of metrics for the probability distributions space. The JSD has shown to be useful in several context both in classical and quantum physics [12–14], in the realm of statistical biology [15,16], and in network theory [17] just to mention a few.

Rényi, by imposing certain properties, introduced the corresponding divergence to his entropy [5], achieving to the quantity

$$\Lambda_\alpha^R\left(\mathcal{P},\mathcal{Q}\right)=\frac{1}{\alpha-1}\log_2\left(\sum_i p_i^\alpha q_i^{1-\alpha}\right) \tag{8}$$

for $\alpha>1$. He interpreted this quantity as "the information of order α obtained if the distribution $\mathcal{P}$ is replaced by the distribution $\mathcal{Q}$". For the Havrda–Charvát entropy, the corresponding divergence is given by [18]

$$\Lambda_q^{HC}\left(\mathcal{P},\mathcal{Q}\right)=\frac{1}{1-q}\sum_i p_i^q\left(p_i^{1-q}-q_i^{1-q}\right) \tag{9}$$

what is a q-average of the change in the information associated to the distributions $\mathcal{P}$ and $\mathcal{Q}$. This last divergence is of Csiszár's type, with $f(x)\equiv\frac{1-x^{1-q}}{1-q}$

Lin [10] among many others, expressed the JSD as a symmetrized version of the KLD. Indeed, the divergence $D_J^S\left(\mathcal{P},\mathcal{Q};\frac{1}{2},\frac{1}{2}\right)$ can be rewritten under the form

$$D_J^S\left(\mathcal{P},\mathcal{Q};\frac{1}{2},\frac{1}{2}\right)=\frac{1}{2}K\left(\mathcal{P},\frac{\mathcal{P}+\mathcal{Q}}{2}\right)+\frac{1}{2}K\left(\mathcal{Q},\frac{\mathcal{P}+\mathcal{Q}}{2}\right) \tag{10}$$

At this point an ambiguity becomes evident. If we use for example the HC entropy in expression (6), the resulting quantity differs from those obtained by using the divergence equation (9) instead of the KLD in (10). In Ref. [19] we studied that differences. The same situation occurs for the Rényi entropy.

The searching for alternative roads to avoid such ambiguities is the main motivation for the present work. Our proposal will provide measures of distinguishability between probability distributions within the framework of generalized entropies. Other generalizations have been proposed and used practically in different contexts (see for example [16,19–24]).

2. Recovering the Bregman divergences

2.1. General proposal

In terms of the Shannon entropy the KLD can be written as

$$K\left(\mathcal{P},\mathcal{Q}\right)=\sum_i p_i\left(\frac{\partial H^S\left(\mathcal{Q}\right)}{\partial q_i}-\frac{\partial H^S\left(\mathcal{P}\right)}{\partial p_i}\right) \tag{11}$$

In vectorial notation it can equivalently be written as:

$$K\left(\mathcal{P},\mathcal{Q}\right)=\boldsymbol{\mathcal{P}}\left(\nabla H^S\left(\mathcal{Q}\right)-\nabla H^S\left(\mathcal{P}\right)\right)^t$$

with $\boldsymbol{\mathcal{P}}=\begin{bmatrix}p_1 & \dots & p_N\end{bmatrix}$ the vector of the probabilities, ∇ the gradient operator, and $\cdot^t$ denoting the transpose operator. In some way, we can think about the KLD as an average of the difference of the entropy gradients, evaluated in each probability distribution. This inspires us a line to define a divergence associated to an arbitrary entropy H^G replacing the Shannon entropy H^S in expression (11) by the generalized entropy H^G. Proceeding in this way, we obtain a quantity $K^G\left(\mathcal{P},\mathcal{Q}\right)$ exhibiting a linear term in δp_i in the Taylor expansion of $K^G\left(\mathcal{P},\mathcal{P}+\delta\mathcal{P}\right)$, which is incompatible with a Csiszár like divergence (obviously, $K^G\left(\mathcal{P},\mathcal{P}\right)=0$ so that the zero-order term is zero). It is easy to conclude that the only entropy, when replaced in (11), that leads to a quantity with quadratic terms as its first non-zero coefficient of the Taylor expansion, is the Shannon entropy. This point can be overcame through a symmetrization, resulting what we will call *generalized divergence*:

$$\Delta^{H^G}\left(\mathcal{P},\mathcal{Q}\right)=\frac{1}{2}\sum_i\left(p_i-q_i\right)\left(\frac{\partial H^G\left(\mathcal{Q}\right)}{\partial q_i}-\frac{\partial H^G\left(\mathcal{P}\right)}{\partial p_i}\right) \tag{12}$$

or in vectorial form:

$$\Delta^{H^G}=\frac{1}{2}\left(\boldsymbol{\mathcal{P}}-\boldsymbol{\mathcal{Q}}\right)\left(\nabla H^S\left(\mathcal{Q}\right)-\nabla H^S\left(\mathcal{P}\right)\right)^t$$

For two "close" distributions, the Taylor expansion of this divergence is given by

$$\Delta^{H^G}\left(\mathcal{P},\mathcal{P}+\delta\mathcal{P}\right)=-\frac{1}{2}\sum_{i,j}\frac{\partial^2 H^G}{\partial p_i\partial p_j}\delta p_i\delta p_j+o\left(\sum_{i,j}\delta p_i\delta p_j\right) \tag{13}$$

In information geometry the coefficients (positive due to the concavity of H^G)

$$g_{ij}(\mathcal{P}) = -\frac{1}{2}\frac{\partial^2 H^G}{\partial p_i \partial p_j}(\mathcal{P})$$

are thought as a Riemannian metric for the manifold (simplex) of the discrete probability distributions [8]. Incidentally, we mention that from the metric $g_{ij}(\mathcal{P})$, the corresponding geodesics could be determined, and from them the geodesic distances could be evaluated [25]. In the case of the Shannon entropy the corresponding Riemannian metric is the Fisher metric and the resulting geodesic distance between two probability distributions $\mathcal{P}$ and $\mathcal{Q}$, is given by:

$$W(\mathcal{P}, \mathcal{Q}) = 2\arccos\left(\sum_i \sqrt{p_i q_i}\right)$$

This distance it is known in physics as the Wootters's distance [26].

Notably expression (12) is closely related to the Bregman divergences [27]. Indeed a Bregman divergence has the expression:

$$d_\psi(\mathcal{P}, \mathcal{Q}) = \psi(\mathcal{P}) - \psi(\mathcal{Q}) - \sum_i (p_i - q_i)\frac{\partial \psi(\mathcal{Q})}{\partial q_i} \tag{14}$$

or in vectorial form:

$$d_\psi(\mathcal{P}, \mathcal{Q}) = \psi(\mathcal{P}) - \psi(\mathcal{Q}) - (\boldsymbol{\mathcal{P}} - \boldsymbol{\mathcal{Q}})\nabla^t \psi(\mathcal{Q})$$

where ψ is a real valued convex function defined on a convex set $\mathcal{S} \subseteq \mathbb{R}^N$ such that ψ is differentiable on the interior of $\mathcal{S}$.

When we insert in (14) the function $\psi(\mathcal{P}) = -H^S(\mathcal{P})$, the Bregman divergence reduces to the KLD. If we substitute $\psi(\mathcal{P}) = -H^G(\mathcal{P})$ in the expression (14) and symmetrize it, quantity (12) is recovered:

$$\Delta^{H^G}(\mathcal{P}, \mathcal{Q}) = \frac{1}{2}\left(d_\psi(\mathcal{P}, \mathcal{Q}) + d_\psi(\mathcal{Q}, \mathcal{P})\right), \tag{15}$$

For the particular case of the Shannon entropy, $\psi(\mathcal{P}) = -H^S(\mathcal{P})$, this last definition leads to the Jeffrey divergence.

For the case of the Rényi entropy, the symmetrized version (12) leads to

$$\Delta^R(\mathcal{P}, \mathcal{Q}) = \frac{\alpha}{2(1-\alpha)}\left(\sum_i \left(\frac{p_i}{q_i} - 1\right) e_i^{(\alpha)}(\mathcal{Q}) + \sum_i \left(\frac{q_i}{p_i} - 1\right) e_i^{(\alpha)}(\mathcal{P})\right) \tag{16}$$

for $0 < \alpha < 1$ and where $A_\alpha(\mathcal{P}) \equiv \sum_i p_i^\alpha$ and the"escort" probability associates with $\mathcal{P}$ are given by $e_i^{(\alpha)}(\mathcal{P}) \equiv \frac{p_i^\alpha}{A_\alpha(\mathcal{P})}$.

In a similar way we achieve to the following expression for the HC entropy

$$\Delta^{HC}(\mathcal{P}, \mathcal{Q}) = \frac{kq}{1-q}\left(\Gamma(\mathcal{P}, \mathcal{Q}) + \Gamma(\mathcal{Q}, \mathcal{P})\right) \tag{17}$$

for $q > 1$ and where $\Gamma(\mathcal{P}, \mathcal{Q}) = A_q(\mathcal{P})\left(-\sum_i \frac{q_i}{p_i} e_i^{(q)}(\mathcal{P}) + 1\right)$.

It turns out that the Bregman divergence satisfies the following well known properties:

[B1] Non-negativity

$$d_\psi(\mathcal{P}, \mathcal{Q}) \geq 0$$

with equality if and only if $\mathcal{P} = \mathcal{Q}$. This is a direct consequence of the convexity of ψ.

[B2] Extensivity

$$d_{\lambda\psi}(\mathcal{P}, \mathcal{Q}) = \lambda d_\psi(\mathcal{P}, \mathcal{Q}); \qquad (\lambda \geq 0) \tag{18}$$

[B3] Convexity $\psi(\mathcal{P}, \mathcal{Q})$ is convex in the first argument, but not necessarily in the second one. This is also a direct consequence of the convexity of ψ.

However $d_\psi(\mathcal{P}, \mathcal{Q})$ it is not symmetric and it does not satisfy the triangle inequality.

The symmetrized version (15) inherits properties [B1] and [B2] and is obviously symmetrical.

Let us now study the divergences Δ^{H^G} for a special class of entropies known as (h, ϕ)-entropies. From these entropies, we will obtain a more general class of divergences that those studied by Burbea and Rao in [8].

2.2. Generalized divergences from (h, ϕ)-entropies

An (h, ϕ)-entropy is defined by [8,9]

$$H_{(h,\phi)}(\mathcal{P}) = h\left(\sum_{i=1}^N \phi(p_i)\right) \tag{19}$$

where the functions $h : \mathbb{R} \mapsto \mathbb{R}$ and $\phi : [0, 1] \mapsto \mathbb{R}$ are such that one, and only one, of these two conditions is fulfilled:

(i) h is increasing and ϕ is concave
(ii) h is decreasing and ϕ is convex

with $\phi(0) = 0$ and $h(\phi(1)) = 0$.

This family of entropies has as particular cases the Rényi and HC entropies. (h, ϕ)-entropies have been used in estimation problems [28] and have been recently extended to realm of quantum physics [29].

It should be noted that for a (h, ϕ)-entropy it is possible to introduce a divergence in the form

$$\mathcal{D}_{(h,\phi)}(\mathcal{P}, \mathcal{Q}) = -h\left(\sum_i q_i \phi\left(\frac{p_i}{q_i}\right)\right) \tag{20}$$

This family of divergence is nothing more than an extension of the Csiszàr class [3,30]. In particular, it allows to include the Rényi divergence. Indeed, if we evaluate this expression for the corresponding functions h and ϕ of the cases of Rényi and HC entropies, one obtains the divergences (8) and (9) respectively.

Let us now take a look at the resulting Bregman divergences when the function ψ is taken to be equal to $-H_{(h,\phi)}$. Not every (h, ϕ)-entropy is adequate to build a Bergman entropy. Indeed, it must be concave, which is not always the case for the Rényi entropy for instance. Assuming that both h and ϕ are of class C^2, $H_{(h,\phi)}$ is also of class C^2. The convexity requirement implies that the Hessian matrix $\mathcal{H}H_{(h,\phi)}$ of $H_{(h,\phi)}$ has elements

$$\left[\mathcal{H}H_{(h,\phi)}\right]_{i,j} = -h''(y)\phi'(p_i)\phi'(p_j) - h'(y)\phi''(p_i)\delta_{i,j} \qquad \text{with} \qquad y = \sum_i \phi(p_i) \tag{21}$$

must be definite positive.

Searching for necessary and sufficient conditions on (h, ϕ) to satisfy the definite positivity of $\mathcal{H}H_{(h,\phi)}$ is not an obvious task. However, one can easily check that if h is concave, then this condition holds. For instance, for the Rényi entropy, this sufficient condition results to be valid for $0 < \alpha < 1$. However, although h in no more concave for $\alpha > 1$, it has been proved that there exists an $\alpha^*(N)$, depending on N such that the Rényi entropy remains concave when $\alpha < \alpha^*(N)$ [31, p. 57] ($\alpha^*(2) = 2$), showing that the condition is only sufficient.

One can easily check that the HC entropy is concave whatever $q > 0$.

3. Weighted generalized divergences

In several contexts it could be useful to assign different relevance to different probability distributions. This is the case for example, as evoked previously, in Bayesian inference. In this section we propose a way to define a generalized divergence between weighted probability distributions. To this aim, we come back to our proposal of divergence equation (12) that turned to be a symmetrization of the Bregman divergence written in terms of general entropies $\psi(\mathcal{P}) = -H^G(\mathcal{P})$ as given by expression (15). The connection between the Jensen–Shannon divergence and the Kullback–Leibler divergence, that is a Bregman divergence associated to the Shannon entropy, suggests an alternative way to generalize the divergence $\Delta^{H^G}(\mathcal{P}, \mathcal{Q})$ when we need to assign different weights to the distributions $\mathcal{P}$ and $\mathcal{Q}$.

Let π_P and π_Q be two non negative numbers such that $\pi_P + \pi_Q = 1$. These numbers can be interpreted as the weights for the distributions $\mathcal{P}$ and $\mathcal{Q}$ respectively. We introduce the weighted average distribution as

$$\mathcal{M} = \pi_P \mathcal{P} + \pi_Q \mathcal{Q}$$

Then we can define a *weighted generalized divergence* as:

$$\Delta^{H^G}\left(\mathcal{P}, \mathcal{Q}; \pi_P, \pi_Q\right) \equiv \pi_P d_\psi(\mathcal{P}, \mathcal{M}) + \pi_Q d_\psi(\mathcal{Q}, \mathcal{M}) \qquad \text{with} \qquad \psi = -H^G \tag{22}$$

After some simple algebra one can check that this measure coincides with those given in expression (7). Definition Eq. (22) provides an unequivocal way of assigning weights in a divergence for any arbitrary (concave) entropy. Let us stress that the natural divergence associate to the entropy to use for this is not that of the Csiszàr class, but that of the Bregman class.

4. Applications

In this section we use the weighted generalized divergences previously introduced to study the stationarity of a time series. We apply an already developed method consisting in a sliding pointer that moves along the register of the time series [32]. For each position of the pointer we evaluate the frequencies of occurrence of the symbols belongings to the alphabet used to map the sequence before and after the pointer, empirical probabilities of occurrence of the symbols. The relative lengths of each subsequence (to the left and to the right of the pointer) is used as the weight of the corresponding frequencies (or estimated probability distributions). Then, we evaluate the divergence (7), (22) at the obtained empirical distributions. The maximum of this quantity, as function of the cursor position, is interpreted as the detection of a position where the distribution of symbols changes. For details see Refs. [13,32].

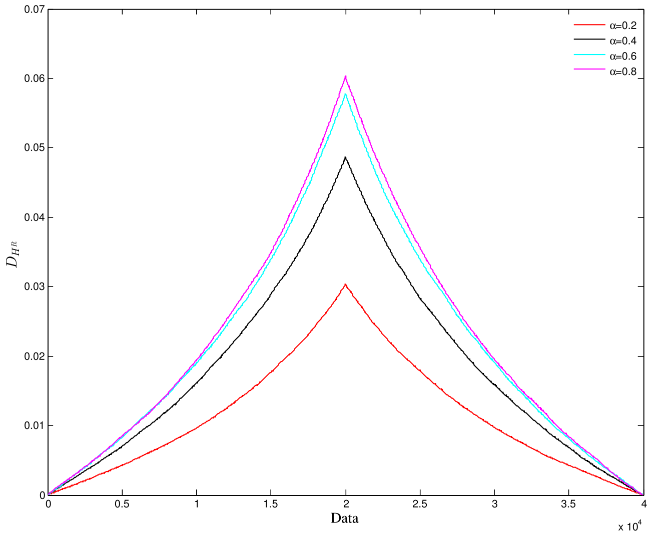

Fig. 1. Averaged weighted generalized divergence corresponding to the Rényi entropy evaluated over the binary sequence above described, as a function of the cursor x. Each curve corresponds to a different value of α in the range $0.2 \leq \alpha \leq 0.8$.

We apply this scheme to a simulated sequence and to a real world sequence. The first one is a binary sequence, which was built form the merging of two binary subsequences generated with different probability distributions for the occurrences of symbols "0" and "1". The second example corresponds to the analysis of an electrocardiogram (ECG) signal, formed by two sub-signals, one belonging to an ECG from a patient suffering an atrial fibrillation (AF) and the other corresponding to a normal sinus rhythmic (NSR). In our examples we use the proposed generalized divergences corresponding to the Rényi and the Tsallis entropies.

4.1. Simulated sequences

We use Monte Carlo simulations to study the behavior of the weighted generalized divergence both in the case of Rényi and HC entropies as a detector of non-stationnarities. In this example, we generated N_e independent sequences composed of two sub-sequences of respective lengths $L_{\mathcal{S}_1}$ and $L_{\mathcal{S}_2} = L - L_{\mathcal{S}_1}$. The sub-sequence of length $L_{\mathcal{S}_k}$ $(k = 1, 2)$ is generated from a probability distribution (vector) $\boldsymbol{\mathcal{P}}_{\mathcal{S}_1} = \begin{bmatrix} s_k & 1 - s_k \end{bmatrix}$ with independent samples. For each realization, we moved a cursor, let us denote x its position, $1 < x < L$, and estimate a distribution $\mathcal{P}_\mathrm{r}(x)$ by the frequencies of 0s and 1's in the subsequences of the first x symbols ("left part"), and a distribution $\mathcal{P}_\mathrm{l}(x)$ by the frequencies of 0s and 1's in the remaining subsequences ("right part"). To take into account the number of samples, i.e., the respective weights of the two subsequences, we define the weights $\pi_\mathrm{r} = \frac{x}{L} = 1 - \pi_\mathrm{l}$. We thus calculated $\Delta^{H^G}(\mathcal{P}_\mathrm{r}(x), \mathcal{P}_\mathrm{l}(x); \pi_\mathrm{r}, \pi_\mathrm{l})$ in both the Rényi case and HC case, for each realization and each x, and compute the averaged divergences $\overline{\Delta}^{H^G}(x) = \left\langle \Delta^{H^G}(\mathcal{P}_\mathrm{r}(x), \mathcal{P}_\mathrm{l}(x); \pi_\mathrm{r}, \pi_\mathrm{l}) \right\rangle$ over the ensemble of the N_e realizations. For the illustration, we have chosen $N_e = 1000$ realization q, sequences of length $L = 40\,000$, with the stationary subsequences of respective lengths $L_{\mathcal{S}_1} = L_{\mathcal{S}_2} = 20\,000$. The distributions of the two subsequences are chosen to be given by $s_1 = \frac{2}{3}$ and $s_2 = \frac{1}{3}$.

Fig. 1 depicts $\overline{\Delta}^{R}(x) = \left\langle \Delta^{R}(\mathcal{P}_\mathrm{r}(x), \mathcal{P}_\mathrm{l}(x); \pi_\mathrm{r}, \pi_l) \right\rangle$ as a function of x using the Rényi entropy, for different values of the parameter $\alpha = 0.2, 0.4, 0.6$ and 0.8. The average reaches the maximum value at the merging point of the two subsequences $x = L_{\mathcal{S}_1}$. Note that the maximum value of the divergence increases w.r.t. α. It should be observed that we evaluate the average in order to get a smooth curve. Individually, for each realization, the divergence Δ^R has obviously a noisy behavior.

Similarly, Fig. 2 depicts $\overline{\Delta}^{HC}(x) = \left\langle \Delta^{HC}(\mathcal{P}_\mathrm{r}(x), \mathcal{P}_\mathrm{l}(x); \pi_\mathrm{r}, \pi_l) \right\rangle$ as a function of x using the HC entropy. The q parameter runs between the values 0.3, 0.5, 0.7 and 1.3 and the data for the Monte Carlo simulations are the same as that used for Fig. 1. Again the average of the divergence reaches its maximum value at the merging point of the two subsequences $x = L_{\mathcal{S}_1}$. Note that, here again, the value of the maximum increases as the values of parameter q increases.

More interesting, Fig. 3 represents the histograms of the position of the cursor when the maximum is reached over the realizations, for various values of q in the Tsallis context (the same behavior occurs for the Rényi case). The curves exhibit a remarkable robustness of the statistics for a wide range of q, the unbiasedness of the maximal position (means precisely equals to 2001) as an estimator of the change of stationarity, and a low variance (standard deviation of 10). This example is a short illustration of how the generalized divergences (re)defined in this paper can be applied in such a context; it merits a deeper study, for instance in terms of analytical study of the bias, variance, confidence interval, curvature of the maximum (for which q has an effective role), etc. Such a study goes beyond the aim of this short paper.

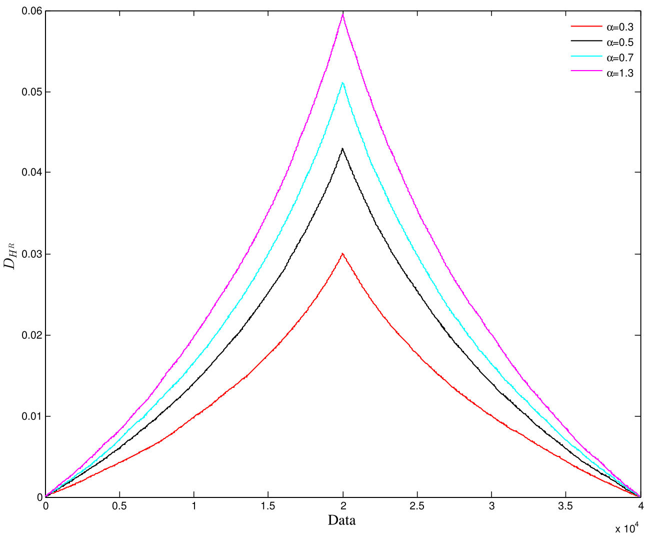

Fig. 2. Averaged weighted generalized divergence corresponding to the HC entropy evaluated over the binary sequence above described, as a function of the cursor *x*.

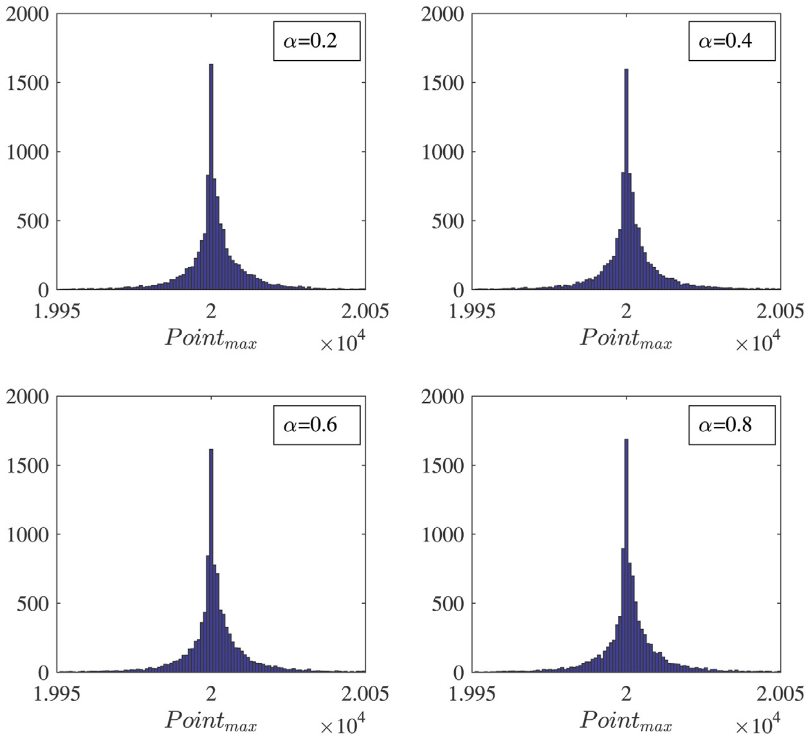

Fig. 3. Histograms of the position of the cursor when the divergence $\Delta^R(\mathcal{P}_r(x), \mathcal{P}_l(x); \pi_r, \pi_l)$ is reached, over the 1000 realizations of the above described process, for various values of the entropic index *q*.

4.2. Atrial fibrillation detection in ECG

Atrial Fibrillation (AF) is very common sustained cardiac arrhythmia, occurring in an important part of the general population [33]. It is associated with significant mortality and morbidity through association of risk of death, stroke, hospitalization, heart failure and coronary artery disease. Despite the enormity of this problem, AF detection remains problematic, because it may be episodic. For these reasons, it is important to develop methods that can detect the difference between AF and Normal Sinus Rhythmics (NSR) using electrocardiogram (ECG) traces.

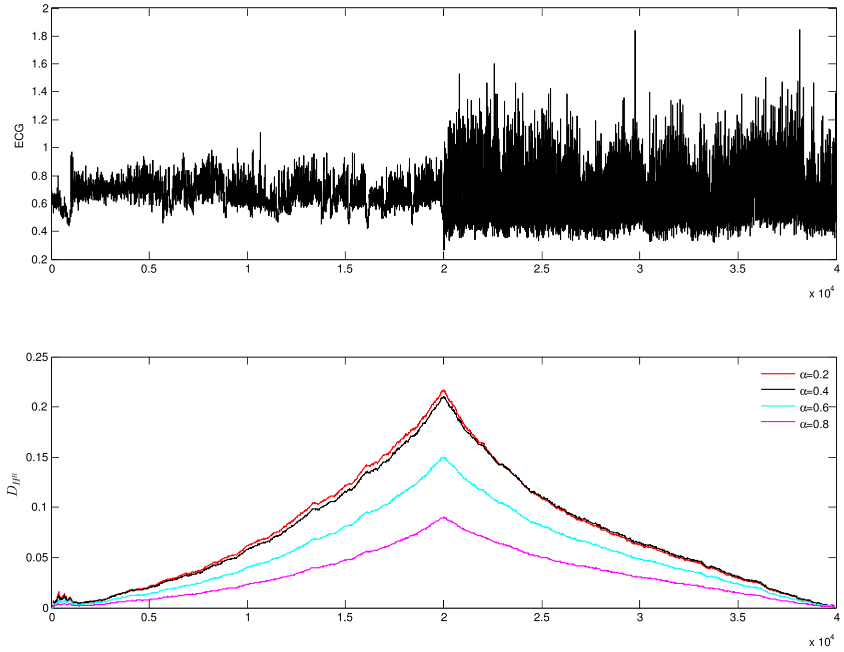

Fig. 4. Weighted Rényi divergence analysis over a normalized ECG signal. The ECG signal is a combination of two sub sequences $\mathcal{S} = \mathcal{S}_1\mathcal{S}_2$ with length $L_{\mathcal{S}_1} = L_{\mathcal{S}_2} = 20\,000$ respectively. The first part of the signal belong to atrial fibrillation trace and the second one is Normal Sinus Rhythmic. The signal was discretized using permutation vector approach with parameter $d = 4$ and $\tau = 1$. The divergence was taken using different values of α indicated in the plots.

We test our analysis method in the problematic of detecting the differences between an AF record from a NSR one. To this aim, we put in a single record two ECG signals, one with AF followed by another one with NSR. These records were taken for the Physionet data bank [34]. The signal were normalized and cut with the same length ($L_{AF} = L_{NSR} = 20\,000$ data point) and joined to form the "complete" signal. The signals were previously mapped in a finite-state sequence using the permutation vector mapping method (see [35] for details). This mapping requires two parameters, usually named d (dimension of the embedding) and τ (delay). In our example we use the values $d = 4$ and $\tau = 1$. After this processing we applied the segmentation method described here above by using the Rényi entropy with different values for α.

Fig. 4 shows the behavior of the divergence Δ^R as a function of the cursor moving along the merged recording. This quantity reaches its maximum value at the exact point where the dynamics of the ECG signal changes. This happens because the empirical probability distribution of the permutation vectors are different when the signal is in AF than NSR. This difference is clearly detected by the divergence for all values of α, but the clearest detections happens for low α.

5. Discussion

The present work can be divided in three parts. In the first one we recover the Bregman divergences from a formal re writing of the analogous of the KLD in the context of a general (convex) entropy. This allows to introduce generalized divergences from generalized entropies out of the line of the class of Csiszàr that have the drawback to achieve a unequivocal definition of such a divergence from an entropy. In second part we investigate the conditions to be satisfied for a (h, ϕ)-entropy in order to define a symmetrized Bregman divergence. Finally we applied the generalized divergences we proposed to the study of simulated and real world time series. Our results showed that these measures could be adequate to discriminate different statistical properties of a generical time series. The existence of free parameters in the definition of some entropies, allows, in principle, to choose the most adequate values according to the problem under study.

The extension of our analysis to the realm of quantum theory is in progress.

Acknowledgments

P.W.L. and L.R. acknowledge the Argentinian agencies SeCyT-UNC and CONICET for financial support. L.R. has a fellowship from CONICET.

References

[1] C.E. Shannon, A mathematical theory of communications, Bell Syst. Tech. J. 27 (1948) 379–423.

[2] T.M. Cover, J.A. Thomas, Elements of Information Theory, Wilet, USA, 2005.
[3] I. Csiszàr, Studia Sci. Math. Hungar. 2 (1967) 299–318.
[4] G.E. Crooks, D.J. Sivak, Stat. Mech: Theory Exp. 2011 (2011) P06003.
[5] A. Rényi, On measures of information and entropy, in: Proceedings of the fourth Berkeley Symposium on Mathematics, Statistics and Probability, 1961, pp. 547–561.
[6] J. Havrda, F. Charvát, Kybernetika 3 (1967) 30–35.
[7] Z. Daróczy, Inf. Control 16 (1970) 36–51.
[8] J. Burbea, R. Rao, IEEE Trans. Inform. Theory 28 (1982) 489–495.
[9] M. Salicrú, M.L. Menéndez, D. Morales, L. Pardo, Comm. Statist. Theory Methods 22 (1993) 2015–2031.
[10] J. Lin, IEEE Trans. Inform. Theory 37 (1991) 145–151.
[11] T. Osán, D. Bussandri, P.W. Lamberti, Physica A 495 (2018) 336–344.
[12] E.H. Feng, G.E. Crooks, Phys. Rev. Lett. 101 (2008) 090602.
[13] D. Mateos, L. Riveaud, P.W. Lamberti, Chaos 27 (2017) 083118.
[14] A. Majtey, D. Prato, P.W. Lamberti, Phys. Rev. A 72 (2005) 052310.
[15] P. Bernaola-Galvan, J.L. Oliver, R.R. Román, Phys. Rev. Lett. 83 (1999) 3336–3339.
[16] M.A. Ré, R.K. Azad, PLoS One 9 (2014) e93532.
[17] M. De Domenico, A. Solé Riviolta, E. Omodei, S. Gómez, A. Arenas, Nature Commun. 6 (2015) 6868.
[18] C. Tsallis, Phys. Rev. E 58 (1998) 1442.
[19] P.W. Lamberti, A.P. Majtey, Physica A 329 (2003) 81–90.
[20] E.G. Altmann, L. Dias, M. Gerlach, J. Stat. Mech. Theory Exp. 2017 (2017) 014002.
[21] A.L. Martin, J.C. Angulo, J. Antolin, S. López Rosa, Physica A 467 (2017) 315–325.
[22] R. Verma, B.D. Sharma, Informatica 37 (2013) 399–409.
[23] F. Nielsen, S. Boltz, IEEE Trans. Inform. Theory 57 (8) (2011) 5455–5466.
[24] F. Nielsen, R. Nock, arXiv:1702.04877v2, 2017.
[25] S. Amari, Methods of Information Geometry, Oxford University Press, 2000.
[26] W.K. Wootters, Phys. Rev. D 23 (1981) 357.
[27] L. Bregman, USSR Comput. Math. Math. Phys. 7 (1967) 200–217.
[28] M.L. Menéndez, J.A. Pardo, M.C. Pardo, Appl. Math. Lett. 11 (1998) 99–104.
[29] G.M. Bosyk, S. Zozor, F. Holik, M. Portesi, P.W. Lamberti, Quantum Inf. Process. 15 (2016) 3393–3420.
[30] G.C. Orsak, B.-P. Paris, IEEE Trans. Inform. Theory 41 (1995) 188–203.
[31] I. Bengtsson, K. Życzkowski, Geometry of Quantum States: An Introduction to Quantum Entanglement, University Press, 2006.
[32] I. Grosse, P. Bernaola-Galvan, P. Carpena, R.R. Roldán, J. Oliver, H.E. Stanley, Phys. Rev. E 65 (2002) 041905.
[33] G.Y.H. Lip, L. Fauchier, S.B. Freedman, I. Van Gelder, A. Natale, C. Gianni, S. Nattel, T. Potpara, M. Rienstra, D.A. Lane, Nat. Rev. Dis. Primers 2 (2016) 16016.
[34] These time series are freely available by PhysioNet (http://www.physionet.org/challenge/).
[35] C. Bandt, B. Pompe, Phys. Rev. Lett. 88 (2002) 124102.