A pragmatic entropy and differential entropy estimator for small datasets

Jugurta Montalvão, Romis Attux, and Daniel Silva,

***Abstract*—A pragmatic approach for entropy estimation is presented, first for discrete variables, then in the form of an extension for handling continuous and/or multivariate ones. It is based on coincidence detection, and its application leads to algorithms with three main attractive features: they are easy to use, can be employed without any *a priori* knowledge concerning source distribution (not even the alphabet cardinality K of discrete sources) and can provide useful estimates even when the number of samples, N, is less than K, for discrete variables, whereas plug-in methods typically demand $N >> K$ for a proper approximation of probability mass functions. Experiments done with both discrete and continuous random variables illustrate the simplicity of use of the proposed method, whereas numerical comparisons to other methods show that, in spite of its simplicity, useful results are yielded.**

***Index Terms*—Entropy through coincidence, Small datasets, Discrete and/or continuous variables, Uncomplicated algorithms.**

I. Introduction

THE entropy of discrete random sources is a pivotal matter in Information Theory (IT). The concept was defined by Shannon and generalized by Rényi's set of parametrized measurements (Rényi, 1961). In both cases, the definition of entropy depends upon the probability associated with each symbol used by the source. Thus, when it comes to entropy estimation, a straightforward first step is to estimate symbol probabilities, whose representations we recognize as common *histograms*. In other words, whenever we need to estimate entropy, a natural approach is to take as many samples as possible to build histograms and then to use these histograms as probability estimators in Shannon's or Rényi's formula. These approaches are known as plug-in methods (Beirlant et al., 1997).

Besides the well-known entropy estimation bias (Miller, 1955), a remarkable issue of plug-in methods is that one must first estimate K probabilities, where K is the number of symbols used by the source. As a consequence, for high values of K (high cardinalities), and/or when some symbols are associated with very low probabilities, a possibly prohibitive number of samples may be necessary to provide reliable estimations, not to mention that K must be known in advance.

Motivated by these practical limitations of plug-in methods, an interesting question can be formulated as: can we get rid of histograms in entropy estimation? Fortunately, the answer is 'yes'. And this answer brings together a series of interesting points of view. Indeed, the key event in any entropy measurement is the coincidence of symbols in a sample. Strictly speaking, any histogram-based estimator relies on coincidence counters, since histogram bins quantify coincidences of each symbol in a stream of symbols. However, using K coincidence detectors can be problematic. For instance, if the number of available samples is less than K, histogram-based estimators are expected to perform badly, since at least one coincidence counter is not incremented at all, thus inducing strong estimator bias and variance. For small data sets and discrete random variables, Bonachela et al. (2008) propose a method to balance estimator bias and variance, along with a very interesting point of view that elegantly links existing methods such as Miller's and Grassberger's to their own approach.

By contrast, an entropy method of estimation through coincidences was proposed by Ma (1980), in a journal paper, and re-explained in a book by the same author (Ma, 1985, Ch. 25) as a 'method (...) in the stage of development' to be used in Statistical Mechanics. This author also discusses an interesting link between IT and Statistical Mechanics, in which he points out that 'In information theory the number of symbols is very small and each symbol is used many times' so that probabilities 'can be accurately determined.' It was certainly the general perception by the time his book was written. Nonetheless, in some hard problems involving blocks of symbols, which may occur in practical domains ranging from multiple-input and multiple-output digital systems to large-scale data mining, even small sets of symbols may lead to problems of entropy estimation with a huge number of states. In other words, nowadays, we conjecture that Ma-like methods can also be attractive for problems belonging to a variety of domains, wherever phenomena with a huge number of reachable states are observed.

This seems to be the motivation behind the method proposed by Nemenman et al. (2002), where entropy estimation through coincidence counting was elegantly revisited in the context of an information-theoretical analysis of neural responses (Nemenman et al., 2004). Not surprisingly, they highlight the benefits of such an approach when the number of samples is smaller than the number of states — the same motivation in Ma's work.

Nemenman (2011) further analyses this estimator previously proposed by himself and collaborators, in 2002. His analysis, to a certain extent, bridges the gap between entropy and differential entropy estimation through their coincidence counting approach, by considering random variables with large cardinalities, and thus coming to the conclusion that the *a*

J. Montalvão is with the Federal University of Sergipe (UFS), São Cristóvão, Brazil e-mail: jmontalvao(at)ufs.br.

R. Attux and D. Silva are with the University of Campinas (UNICAMP), São Paulo, Brazil. D. Silva is also with the University of Brasília (UnB)

priori knowledge of the cardinality of the alphabet size is not necessary. It is noteworthy that it allows for the estimation of differential entropies, where cardinalities tends to infinity. Unfortunately, in spite of this open possibility, the analysed method was not adapted to continuous random variables. By following the same path, in (Montalvão et al., 2012) we briefly proposed a simpler method (for discrete variables only) which can be used without any knowledge of the cardinality of the alphabet size, and is simple enough to be easily employed even by experimenters unfamiliar with the theoretical bases of statistical estimation.

In this paper, we extend this method toward continuous multivariate random sources, keeping, however, simplicity of use as a *leitmotiv*, along with the method's suitability of use with small datasets. In order to properly introduce this method extension, in Section II, we first recall the method proposed in (Montalvão et al., 2012), along with some new theoretical explanations of an important approximation used there and an analysis of the computational burden associated with it. Experimental results with discrete random variables are presented in Section III. The method generalization, which, as already stated, is the main novelty brought forward in this work, is presented in Section IV, whereas experimental results with continuous random variables are presented in Section V. A section devoted to the conclusions and to a final discussion closes the paper.

II. Proposed method for entropy estimation

Instead of counting coincidences of each symbol, as in histogram-based approaches, we address entropy estimation by detecting any coincidence of symbols. For memoryless random sources, this unconstrained coincidence detection is closely related to the classical 'Birthday Problem', presented in textbooks of probability (Papoulis, 2002). By generalizing this problem, let K be the number of equiprobable symbols – if they are independently drawn from this source, the probability of repeating one or more symbols by the x-th sample is given by:

$$F_X(x;K) = 1 - \frac{K(K-1)(K-2)\dots(K-x+1)}{K^x} \quad (1)$$

where $x \in \{1,2,3,\dots,K+1\}$, K plays the role of a parameter for this Cumulative Distribution Function (CDF), and the probability of a first coincidence precisely at the x-th sample is given by $f_X(x;K) = F_X(x;K) - F_X(x-1;K)$. Therefore, we can estimate the average number of samples drawn from the source until a first coincidence occurs as:

$$D(K) = \sum_{x=1}^{K+1} x f_X(x;K) \quad (2)$$

which clearly depends on K. For instance, in the Birthday Problem itself, where $K = 365$ days, on average, we shall expect one birthday coincidence roughly every 24 independently consulted subjects. Figure 1 graphically presents D as a function of K, from $K = 2$ to $K = 2000$.

Now, by considering the inverse function, $g(D) = K$ (i.e. by exchanging axis in Figure 1), we observe a striking quadratic functional dependence, which can be suitably approximated as in Eq. 3.

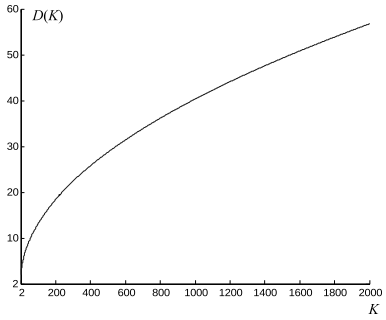

Fig. 1. Average number of symbols, D, drawn from a white source of K equiprobable symbols until a first coincidence occurs.

$$K \approx g(D) = aD^2 + bD + c \quad (3)$$

Indeed, through squared error minimization, we obtain $a = 0.6366$, $b = -0.8493$ and $c = 0.1272$, which yields a Mean Squared Error between K and $g(D)$ of about 10^{-6}, inside the interval $D(1) = 2$ to $D(2000) \approx 56.7$. This polynomial approximation is a key aspect of the method proposed here.

On the other hand, in Shannon's definition of entropy, as well as in Rényi's generalization, whenever all the K symbols of a memoryless random source are equiprobable, the source entropy, in bits, equals $\log_2(K)$. In other words, the entropy, H, of a given non-equiprobable source informs us that there is an "equivalent" source of 2^H equiprobable symbols. By keeping this in mind, we now may consider again the non-equiprobable source of symbols. Clearly, we still may empirically estimate $\hat{D}$ by sequentially observing symbols and averaging the number of symbols until a coincidence occurs, as in Figure 2. Although the sources are no longer equiprobable, the measured $\hat{D}$ does still point out a hypothetical equiprobable source of $\hat{K}$ symbols that could provoke the very same average interval. Thus, we conjecture that

Conjecture C_0**:** A source of symbols (not necessarily equiprobable) that provokes the same average interval D as an equiprobable source of cardinality K has the same entropy $H = \log_2 K$ bits.

As a result, the proposed pragmatic method for entropy estimation can be summarized in three steps:

1 Estimate D by sequential observation of symbols, as illustrated in Figure 2, thus obtaining a $\hat{D}$ that can be gradually refined.
2 Compute $\hat{K}(\hat{D}) = a\hat{D}^2 + b\hat{D} + c$, with $a = 0.6366$, $b = -0.8493$ and $c = 0.1272$.
3 Estimate the entropy of the memoryless source, in bits, as $\hat{H} = \log_2(\hat{K})$.

A. On the polynomial approximation $K \approx aD^2 + bD + c$

According to Eq. 2, we define a random variable X whose Probability Mass Function (PMF) is $f_X(x;K)$, and D is the expectation of X, i.e. $D(K) = E_X\{X\}$. The precise value of D is obtained after the calculation of F_X, as in Eq. 1, from which we obtain the PMF f_X, and finally the average

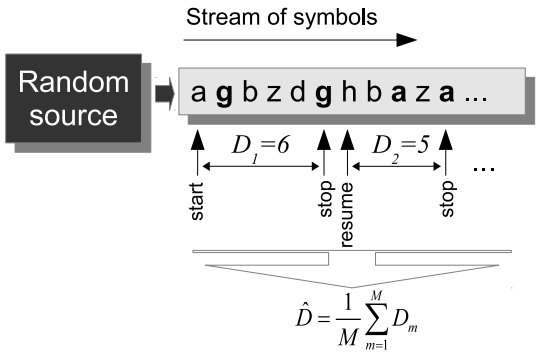

Fig. 2. Incremental estimation of the averaged number of symbols until coincidence detection.

X, as in Eq. 2. This indirect procedure does not show why the second degree polynomial is predominant insofar as the functional dependence of K on D, approximated by Eq. 3, is concerned. In this Section, this quadratic character is analyzed in more detail.

Figure 3 presents visual examples of $F_X(x;K)$, along with their corresponding probability mass functions, $f_X(x;K)$, for $K = 200, 400, 800$ and 1600. Still in Figure 3, the corresponding average values of X are pointed out, for each value of K, lying close to the coordinates of the corresponding peaks of $f_X(x;K)$, which, in turn, correspond to spots of high slopes for $F_X(x;K)$. Given the sigmoidal shape of $F_X(x;K)$, it is expected that this high slope interval is to be found at $F_X \approx 0.5$. More precisely, because of the skewness of $f_X(x;K)$, a better approximation is given by $F_X(E_X\{X\};K) \approx 0.55$ (or some value between 0.5 and 0.6).

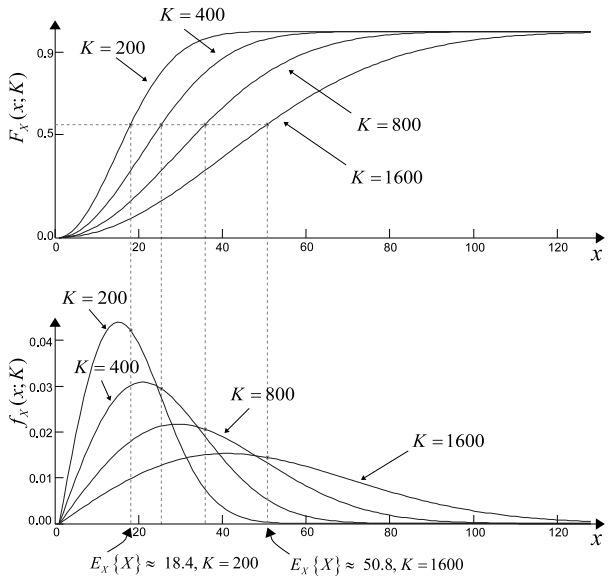

Fig. 3. Visual examples of probability distributions for $K = 200, 400, 800$ and 1600.

Now, because $F_X(E_X\{X\};K) \approx 0.55$, from Equation 1, it follows that

$$\frac{(K-1)(K-2)\dots(K-E_X\{X\}+1)}{K^{E_X\{X\}}} \approx 0.45$$

or

$$(1-1/K)(1-2/K)\dots(1-(E_X\{X\}-1)/K) \approx 0.45 \quad (4)$$

Another useful approximation to be used here is $(1-n/K) \approx \exp(-n/K)$ for $|n| << K$. By assuming that source cardinality, K, is high enough to allow $E_X\{X\} << K$ to hold (please notice that these are the cases our method is designed for), we can apply this approximation to each factor on the left of Equation 4 to obtain

$$\exp(-1/K)\exp(-2/K)\dots\exp(-(E_X\{X\}-1)/K) \approx 0.45$$

which simplifies to

$$\exp(-E_X\{X\}(E_X\{X\}-1)/(2K)) \approx 0.45$$

By taking the logarithm of both sides of the former expression, it follows that

$$-E_X\{X\}(E_X\{X\}-1)/(2K) \approx \ln(0.45)$$

that can be finally rearranged as in Eq. 5 to highlight the quadratic dependency of K on $E_X\{X\}$.

$$K \approx \alpha E_X\{X\}^2 + \beta E_X\{X\} \quad (5)$$

where $\alpha = \frac{-1}{2\ln(0.45)} \approx 0.6261$ and $\beta = -\alpha$.

Therefore, the quadratic dependency empirically adjusted in Eq. 3 is finally justified by Eq. 5. Notwithstanding, the two polynomials clearly have discrepant coefficients, and we propose that Eq. 5 should be regarded only as a confirmation of the quadratic dependency empirically noticed, whereas Eq. 3 is the one we should use within the proposed method, because of its better approximation of K.

B. Computational burden of the method

Approximation 5 can be rewritten as

$$E_X\{X\}^2 - E_X\{X\} + 2K\ln(0.45) \approx 0 \quad (6)$$

whose roots are given by: $(1/2)(1 \pm \sqrt{1-8K\ln(0.45)})$. Only the positive root is a valid estimate of $E_X\{X\}$. Therefore

$$E_X\{X\} \approx (1/2)(1+\sqrt{1+6.4K}) \quad (7)$$

We recall that, in average, we should expect one coincidence detection every $E_X\{X\}$ symbols. Moreover, according to Eq. 7, for $K >> 2$ the value of $E_X\{X\}$ is almost proportional to $\sqrt{K}$. It is worth noting that this is the worst case, for K equiprobable symbols, being the value smaller for non-equiprobable distributions.

Thus, after every detection, all past symbols are discarded and a new subsequence of about $\sqrt{K}$ is observed until new detection, and so forth. Consequently, in average, a set of N sequential observations (N symbols) is to be split into $B = N/\sqrt{K}$ sub-sequences.

Inside each sub-sequence, the ith new symbol is compared to all the $i-1$ symbols observed after the last coincidence detection, yielding $i(i-1)/2$ pairwise comparisons. Since each subsequence is expected to have $D \approx \sqrt{K}$ symbols, in average, then it yields a computational burden of about $\sqrt{K}(\sqrt{K}-1)/2$ comparisons per sub-sequence, and a total of $B\sqrt{K}(\sqrt{K}-1)/2$ comparisons, which simplifies to

$N(\sqrt{K}-1)/2$. Therefore, this is the expected computational burden of the proposed method for the case of a sequence of N observations drawn from a random source of K equiprobable symbols.

This is less than the NK comparisons necessary to obtain histograms in plug-in methods. Moreover, unlike plug-in methods that demand $N >> K$ for a proper approximation of the PMF, the proposed method can be used even when $N < K$. That is to say that, besides its lower computational cost, it can be further lowered through the use of smaller sets of observations (i.e. smaller values on N).

III. Experiments with discrete variables

In agreement with the motivation underlying the pragmatic approach chosen for this work, we will first deal with the emblematic source of 365 equiprobable symbols (from the birthday problem), and we measure its bias in several scenarios. In this case, the known source entropy is $H = \log_2(365) = 8.51$ bits, for the second column in Table I, where this (equiprobable) source was simulated and $\hat{D}$ was obtained through the observation of N sequencial symbols (with at least one coincidence). Then, we applied the proposed method and calculated the relative bias $(\hat{H}-H)/H$ for 10^4 independent trials. Similarly, in the third column, we present the average relative bias for sources whose probability distributions were randomly generated (thus yielding $H \leq \log_2(365)$).

TABLE I
Average estimation of relative biases $\left(\frac{\hat{H}-H}{H}\right)$*, for memoryless sources of* $K = 365$ *symbols, after* N *sequential observations.*

N	Uniform distribution	Random distributions
50	-0.0427	-0.0459
100	-0.0132	-0.0280
1000	-0.0013	-0.0159

It is worth noting that, even for only 50 symbols (i.e., much less than the cardinality of the set, $K = 365$), the average absolute bias is not greater than 5% of the actual entropy of the equiprobable source. Moreover, it is also noteworthy that, for stationary processes, the value of $\hat{D}$ can be iteratively improved, even when K is not known.

To provide a comparative perspective for the proposed method, we address the results presented by Nemenman et al. (2002), in which the authors consider the Dirichlet family of *priors*. In the middle of the entropy range, typical distributions from these priors are "sparse", as illustrated in Figure 4. We consider here the distribution family with cardinality $K = 1000$ and $H = 5.16$ bits.

Table II shows the average estimation biases resulting from the proposed method, for randomly generated "sparse" distribution with $K = 1000$ and $H \approx 5.2$ bits. Again, each row corresponds to 10^4 independent experiments. The relative bias, for these sparse distributions, is much higher than what was found in Table I, even as compared to the worst case of the non-equiprobable sources.

On the other hand, because we know that in all cases the bias results from the underestimation of $\hat{D}$ by about 1 to 2, a

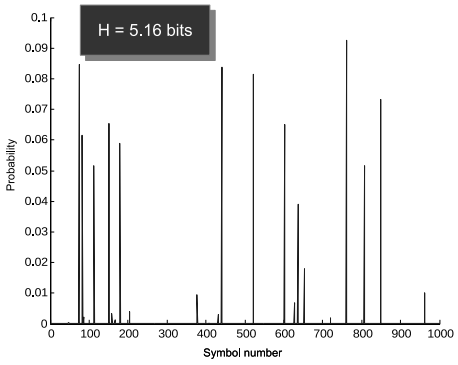

Fig. 4. A typical "sparse" from the Dirichlet family of priors, $K = 1000$.

TABLE II
Average estimation biases for memoryless sources with "sparse" distribution from the Dirichlet family of priors.

N	Relative bias: $(\hat{H}-H)/H$	Standard deviation
10	-0.2767	0.237
30	-0.1918	0.131
100	-0.1829	0.073
300	-0.1795	0.047
1000	-0.1782	0.033
3000	-0.1776	0.029

pragmatic bias compensation procedure is the replacement of $\hat{D}$ with $\hat{D}+1$ in the first step of the algorithm. Indeed, with this compensation, relative biases in the last row of Table I worsen to $(\hat{H}-H)/H \approx 0.01$, for the uniform distribution, but it improves to ≈ 0.0003, for the non-uniform one. Likewise, in the last row in Table II, the relative bias is reduced to ≈ -0.0864, which is a result comparable to that presented in (Nemenman et al., 2002) for the same problem.

IV. Method generalization for differential entropy estimation

The entropy of continuous random variables diverges to infinity, and the usual approach to show this divergence (Cover & Thomas , 1991) is that of partitioning the variable domain into regular cells of size Δ, and to reduce the size of these cells while calculating the corresponding entropy, $H(\Delta)$. Under the requirement that Probability Density Functions (PDF) are continuous inside the cells, if they are sufficiently small, the probability density inside each cell tends to be uniform, so that the corresponding entropy increases by 1 bit whenever the cell size is divided by 2. Therefore, entropy tends to be a linear function of $\delta = \log_2(1/\Delta)$, for sufficiently small values of Δ. Notice that Δ may stand for bin width for 1D variables, bin area for 2D variables, bin volume for 3D variables and so forth.

Accordingly, the differential entropy, h, can be defined as the difference between $H(\delta)$ and the reference $R(\delta) = \delta$ that linearly increases 1 bit per unitary increment of δ. Consequently, h remains almost constant for small enough values of Δ, as illustrated in Figure 5.

By assuming that Δ plays the role of a sample neighbourhood, *coincidence* can be (re)defined for continuously

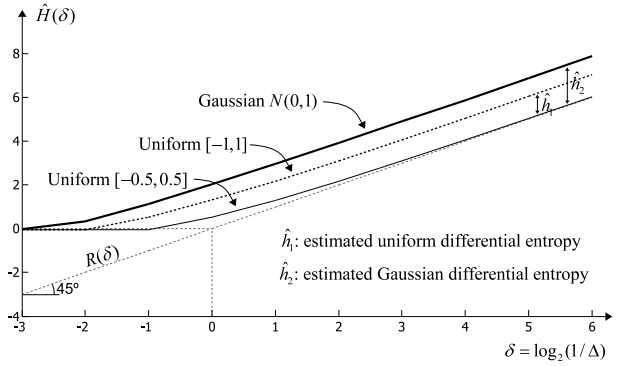

Fig. 5. Graphic explanation of differential entropy.

valued variables as being *a signal sample falling inside the neighbourhood of another sample*. Consequently, for small values of Δ, and a hypothetical continuous but finite variable domain of size A, we can roughly arrange up to $K = A/\Delta$ samples without falling inside each other's neighbourhood.

Through this definition, we are able to estimate H as the entropy of a continuous-valued random variable quantized in about $K = A/\Delta$ disjoint cells. Moreover, for a sufficiently small Δ, we can further estimate the value of h by comparing it to the the reference line $R(\delta)$. Thus, our method can be extended to continuous random variables as:

1 Arbitrarily set a small sample neighbourhood Δ (see Subsection IV-A).
2 Estimate D by sequentially observing samples and detecting coincidences, as illustrated in Figure 6, thus obtaining a $\hat{D}$ that can be gradually refined.
3 Compute $\hat{K}(\hat{D}) = a\hat{D}^2 + b\hat{D} + c$, with $a = 0.6366$, $b = -0.8493$ and $c = 0.1272$.
4 Estimate the entropy of the quantized variable, for the chosen Δ, as $\hat{H}(\Delta) = \log_2(\hat{K})$.
5 Estimate the differential entropy as $\hat{h} = \hat{H}(\Delta) - R(\delta)$ or, equivalently,

$$\hat{h} = \hat{H}(\Delta) + \log_2 \Delta \tag{8}$$

Figure 5 illustrates the results of this method for three 1D random variables, two of them being uniformly distributed, and one normally (Gaussian) distributed. In that Figure, it can be noticed that both differential entropy estimates, $\hat{h}_1$ and $\hat{h}_2$, approach $\log_2(2) = 1$ and $0.5\log_2(2\pi e) \approx 2.047$, respectively, whereas the differential entropy of the uniform PDF with a unitary domain (i.e uniform PDF from -0.5 to 0.5) approaches the reference $R(\delta)$, thus yielding a null differential entropy $\hat{h}_0$ (not shown in the Figure) as expected (Cover & Thomas , 1991).

Another interesting point to be highlighted is that, in spite of the unbounded character of the domain associated with the Gaussian variable, $\hat{h}_2$ also converges to a constant, indicating that there is a uniform distribution with domain size equal to $A = 2^h$ which yields the same average interval $\hat{D}$ until a coincidence occurs. This is analogous to the reasoning that supports *Conjecture* C_0, in Section II.

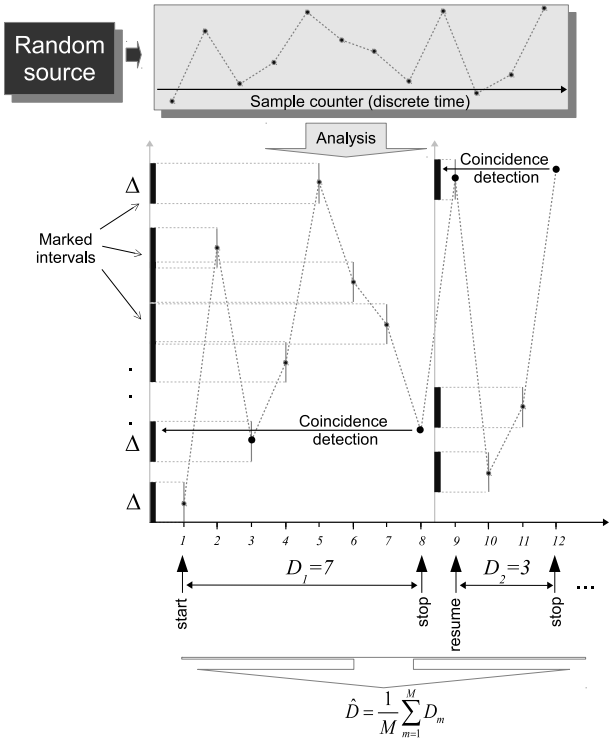

Fig. 6. Incremental estimation of the average number of symbols until coincidence detection, for continuously valued samples. Numerical samples are sequentially compared to all other samples, from the latest "start" or "resume" position, until a new sample falls inside a region already marked by a former sample. When it does occur for a pre-defined region size (Δ), a coincidence is detected, the corresponding delay (D) is recorded, and this process is resumed from next sample.

A. Setting and testing a coincidence neighbourhood

Unfortunately, unlike the procedure presented in Section II, the extension of the method toward differential entropy imposes the choice of an arbitrary value for Δ. We would set it to be as small as possible. But in most realistic scenarios where a limited (possibly small) number N of samples is available, a Δ that is too small may reduce too much the number of coincidences, even to zero. On the other hand, as the proposed method is based on the averaging of intervals, we should choose a Δ corresponding to a tradeoff between the requirement of an approximately constant PDF inside any cell of size Δ, and a number of coincidence within N samples that is not too small. As a rule of thumb, we can set a targeted interval, D_0, which yields the following rough Δ value choice.

From Equation 5, we know that D_0 is approximately given by

$$D_0 \approx \sqrt{K/\alpha} \tag{9}$$

On the other hand, we know that $K \to A/\Delta$ as $\Delta \to 0$ for any uniformly distributed signal over a unitary hypercube of edge $A^{1/L}$, where L stands for the space dimension. Therefore, by denoting the corresponding signal standard deviation in each direction as σ_S, whose value is $\sigma_S = \frac{\sqrt[L]{A}}{2\sqrt{3}}$ for uniform marginals, we can use $A = (2\sqrt{3}\sigma_S)^L$ to approximate

$$K \approx \frac{(2\sqrt{3}\sigma_S)^L}{\Delta} \tag{10}$$

Please note that the induced cardinality, K, diverges to infinity as Δ goes to zero, as well as $H(\delta)$, but the difference between $H(\delta)$ and $R(\delta)$ converges to a finite value.
By applying Eq. 10 to Eq. 9, we obtain $D_0 \approx ((2\sqrt{3}\sigma_S)^L/\alpha\Delta)^{1/2}$ which, after some approximations, yields

$$\Delta \approx \frac{1.6(3.5\sigma_S)^L}{D_0^2} \tag{11}$$

For instance, for an approximated rate of one coincidence every 20 samples and a normalized variance, $\sigma_S^2 = 1$, one should set $\Delta = \frac{1.6(3.5)}{20^2} \approx 0.014$, whereas for a 3D signal with normalized variance in each dimension, at the same coincidence rate, we should set $\Delta = \frac{1.6(3.5)^3}{20^2} \approx 0.171$.

We recall that, in L-dimensional observation spaces, Δ can be an interval, an area, a volume or a hypervolume around any observed sample (a sample neighbourhood). Therefore, we may consider a segment, a square, a cube or a hypercube (i.e. figures with all edges the same length $\sqrt[L]{\Delta}$) so that coincidence detection can be done by simple comparison of distances between samples in each dimension to a threshold of half the edge length.

Nevertheless, any arbitrary Δ, including the one proposed in Eq. 11, should not be used without a test. Indeed, any value of Δ that is small enough must lie in the asymptotically linear part of $H(\delta)$ (as illustrated in Figure 5). Therefore, one can test whether a given Δ is small enough by comparing $\hat{H}(\Delta)$ to $\hat{H}(2\Delta)$, whose difference should be 1 bit. Alternatively, an equivalent test can be done with a much smaller perturbation of Δ, as follows:

$$\hat{H}(\Delta) - \hat{H}(1.05\Delta) \stackrel{?}{\approx} 0.07\ bit \tag{12}$$

If not, then Δ is not small enough to allow a proper use of Eq. 8.

V. Experiments with continuous variables

In this Section, the use of the proposed method with continuous random variables in 1D, 2D and 3D is illustrated through the definition of coincidence neighbourhood as an interval of length Δ, a square of area Δ and a cube of volume Δ, respectively. In the 1D experiment, we set Δ according to Subsection IV-A, with an approximated rate of one coincidence every 30 samples. Four random variables were used, namely: uniform from -1 to +1, Gaussian with unitary variance, exponential with unitary parameter and Laplacian with unitary parameter, respectively yielding the following differential entropies: 1, 2.047, 1.4427 and 2.4427 bits. Table III shows the experimental results in terms of absolute bias and standard variations.

It is worth noticing that these results do not depend on the actual targeted entropy. For instance, a uniform distribution inside the interval $[-512, +512]$ has a differential entropy of $\log_2(1024) \approx 10$ bits. Therefore, by using 100 samples to estimate this entropy with the proposed method, we would expect an estimate lowered by a bias of about -0.13, in average, around which we expect an estimation error with standard deviation of about 0.9. Similarly, a Laplacian distribution with parameter $\lambda = 200$ also has a differential entropy near 10 bits (because $\log_2(2e\lambda) \approx 10$ bits). Therefore, by using only 100 samples to estimate this entropy with the proposed method, we would expect an estimate lowered by a bias of about -0.55, in average, around which we expect an estimation standard deviation of about 0.9.

Similarly to the results presented in Section III, with discrete variables, a bias is also clearly noticed. It is known that histogram-based estimators are biased because symmetrically distributed probability estimates are nonlinearly transformed by a logarithm function, thus biasing the estimated entropy toward lower values (Miller, 1955; Beirlant et al., 1997) . In the proposed method, we also apply a logarithm to random estimates of K, which is indeed the main cause of the noticed bias. However, the random estimate of K polynomially depends on random estimates of D, which in turn are not symmetrically distributed, as shown in Figure 3. As a result, bias analysis is a more difficult matter in this method than in histogram-based ones, and both positive or negative biases are possible.

By contrast, we empirically noticed that, as the joint effect of the logarithmic transformation and the skewness of D estimate distribution depends on its average value, bias can be partially compensated through a careful choice of the targeted D_0 (through the choice of Δ, as in Eq. 11). More specifically, we noticed that by setting D_0 to about 30, a good bias compensation was obtained for uniform distributions.

To provide a brief comparison, it is known that the systematic bias in histogram-based estimators can be approximated through the popular $O(1/N)$ bias compensation method proposed by Miller (1955):

$$\hat{H} \approx H - \frac{K-1}{2N}$$

Accordingly, if we use, for instance, a histogram with $K = 50$ bins and $N = 1000$ samples (as in the last line of Table III), we should expect a bias of about 0.024 for the uniform PDF, which is much higher than the bias presented in Table III, for the same number of samples.

Despite the noticed bias, for high targeted values of h, the method provides meaningful estimates even with only 100 samples for both PDFs, though it is clearly better for uniform ones.

As suggested by Beirlant et al. (1997) and references therein, the number of samples needed for good estimates increases rapidly with the dimension of multivariate densities. By doing similar experiments with the same random variables over 2D and 3D domains (through multivariate samples), we obtained the results presented in Table IV, that illustrate that both absolute biases and variances, as expected, tend to be reduced as the number of samples increases, whereas it is degraded by the increasing of the variable domain dimension. Targeted D_0 was set to 30 through all experiments.

As stated in the introduction of this paper, our main inspiration is the method by Ma (1980), based on the counting of state (physical) system configuration coincidences, in the context of Statistical Mechanics. To allow for comparisons between methods, we replace Ma's definition of coincidence, in terms of particle trajectory of motion (Ma, 1985, Ch. 25), with our

TABLE III
Average estimation of absolute biases $\hat{h} - h$, for memoryless sources of continuously valued samples. Each row corresponds to 10^4 independent trials of N sequential observations (trials without coincidences are discarded).

N	Uniform [-1,+1]		Gaussian N(0,1)		Exponential $\lambda = 1$		Laplace $\lambda = 1$	
	bias	std. dev.	bias	std. dev.	bias	std. dev.	bias	std. dev.
50	-0.592	1.27	-0.862	1.3	-0.737	1.2	-0.928	1.3
100	-0.127	0.91	-0.360	0.9	-0.514	0.8	-0.548	0.9
500	-0.018	0.37	-0.234	0.4	-0.413	0.3	-0.433	0.4
1000	-0.003	0.26	-0.215	0.3	-0.408	0.2	-0.413	0.3

TABLE IV
Average estimation of absolute biases $\hat{h} - h$, for memoryless sources of 2D and 3D random vectors. Each row corresponds to 10^4 independent trials of N sequential observations (trials without coincidences are discarded).

N	Uniform (2D)		Gaussian (2D)		Exponential (2D)		Laplace (2D)	
	bias	std. dev.	bias	std. dev.	bias	std. dev.	bias	std. dev.
100	-0.09	0.9	-0.57	0.9	-0.78	0.7	-0.94	0.8
1000	0.03	0.3	-0.41	0.3	-0.69	0.2	-0.81	0.2
N	Uniform (3D)		Gaussian (3D)		Exponential (3D)		Laplace (3D)	
	bias	std. dev.	bias	std. dev.	bias	std. dev.	bias	std. dev.
100	-0.01	0.9	-0.73	0.9	-0.74	0.6	-1.26	0.8
1000	0.14	0.3	-0.60	0.3	-0.73	0.2	-1.24	0.2

generic definition, as presented in Section IV. Furthermore, for the reader's convenience, we now present the Ma's method in terms of our notation, as follows:

1 Arbitrarily set a small sample neighbourhood Δ (see Subsection IV-A).
2 Compare the two samples in every distinct pair of signal samples, and count the N_c detected coincidences, out of $N_t = N(N-1)/2$ comparisons.
3 Compute the estimated set cardinality as $\hat{K}_{Ma} = N_t/N_c$.
4 Estimate the entropy of the quantized variable, for the chosen Δ, as $\hat{H}_{Ma}(\Delta) = \log_2(\hat{K}_{Ma})$.
5 Estimate the differential entropy as $\hat{h}_{Ma} = \hat{H}_{Ma}(\Delta) - R(\delta)$ or, equivalently, $\hat{h}_{Ma} = \hat{H}_{Ma}(\Delta) + \log_2 \Delta$

By using Ma's method through simulation scenarios equivalent to those of Table IV, we obtained the results presented in Table V.

It is clear that the Ma's method is consistently superior than ours, in terms of variance, which is counterbalanced by an increase in the associated computational burden. More precisely, Ma's method reduces the estimator variance by first gathering all N samples, and then comparing all $N(N-1)/2$ pair of samples, whereas our sequential method compares about $N(D_0 - 1)/2$ pairs instead (see explanation in Subsection II-B). For instance, with $N = 1000$ and $D_0 \approx 30$, Ma's method is expected to demand 499500 samples comparisons whereas our method demands about 14500 comparisons instead. On the other hand, since samples are assumed to be independent, by permutating the N signal samples and applying our method many times, the estimator variance could also be reduced, but it would prevent the desirable sequential aspect of our method (for all N samples would be available before permutation).

VI. Discussion and conclusion

A pragmatic approach for entropy estimation was proposed for both discrete and continuous random sources. This approach can be used when the amount of available data is small, even less than the cardinality of the symbol set for discrete sources, and it chiefly relies on the averaging of intervals until coincidences occur and on a polynomial approximation of a random variable expectation. As a result, the method is very simple to use, thus being potentially useful in applications where experiments are hard/expensive to be reproduced. For discrete sources, the *a priori* knowledge of set cardinality is not even required. The only requirement is that samples be independently drawn from the source (no memory).

The method extension to continuous variables is naturally obtained through the (re)definition of a 'coincidence' in continuous domains. Indeed, some suggestions concerning intervals (sample neighbourhoods) for coincidence detection are presented. Nonetheless, because the only adaptation necessary to transit from discrete to continuous is a suitable definition of coincidences, it can be further explored to allow for joint entropy estimation of discrete and continuous variables that are mixed up. This possibility can be particularly useful in data mining.

In this short text, estimation bias is only illustrated through experimental results, along with practical advices to reduce it. A theoretical analysis of it is planned for future work. However, despite the noticed bias, it was also illustrated that the proposed method provides meaningful estimates even from small datasets.

The presented approach is closely related to the method proposed by Ma (1980), in the context of Statistical Mechanics, as well as to the more recent method proposed by Nemenman et al. (2002), in the context of information analysis of neural (biological) responses. Ma's method is the main source of inspiration for the new method proposed here,

TABLE V
Average Ma's method (Ma, 1980) estimation of absolute biases $\hat{h} - h$, for memoryless sources of 2D and 3D random vectors.

N	Uniform (2D)		Gaussian (2D)		Exponential (2D)		Laplace (2D)	
	bias	std. dev.	bias	std. dev.	bias	std. dev.	bias	std. dev.
100	0.13	0.6	-0.33	0.6	-0.72	0.4	-0.79	0.5
1000	0.03	0.1	-0.44	0.1	-0.78	0.1	-0.88	0.1
N	Uniform (3D)		Gaussian (3D)		Exponential (3D)		Laplace (3D)	
	bias	std. dev.	bias	std. dev.	bias	std. dev.	bias	std. dev.
100	0.23	0.6	-0.54	0.6	-0.84	0.4	-1.18	0.5
1000	0.13	0.1	-0.65	0.1	-0.88	0.1	-1.27	0.1

however, different from our approach, Ma's method counts all coincidences in the data set by testing all available data pairs, whereas we detect coincidences sequentially and completely reset our coincidence search after every new detection, thus considerably reducing the computational burden, and allowing estimation through time (online). As for the method extension to handle continuous variables, there is an interesting link between it and the K-Nearest Neighbours (K-NN) based method proposed by Kraskov et al. (2004) for mutual information estimation. Indeed, we believe that joint coincidences (or synchronized coincidences) are an interesting and powerful approach for easy and reliable mutual information estimators, in the spirit of the discussed framework.

ACKNOWLEDGMENT

This work is supported by the CNPq and FAPESP (grant 2013/11769-3).

REFERENCES

Beirlant, J., Dudewicz, E. J., Gyorfi, L., Van Der Meulen, E. C., 1997. Nonparametric entropy estimation: an overview. International Journal of Mathematical and Statistics Sciences. 6, 17–39.

Bonachela, J. A., Hinrichsen, H., Muñoz, M. A., 2008. Entropy estimates of small data sets. J. Phys. A: Math. Theor. 41, 1–9.

Cover, T. M., Thomas, J. A., 1991. Elements of Information Theory. Wiley.

Kraskov, A., Stögbauer, H., Grassberger, P., 2004.Estimating mutual information. Phys. Rev. E. 69, No. 6, 066138-1–066138-16.

Ma, S.-K., 1980. Calculation of Entropy from Data of Motion. Journal of Statistieal Physics, Vol. 26, No. 2, 221–240.

Ma, S.-K., 1985. Statistical Mechanics. World Scientific Publishing Co. Pte. Ltd.

Miller, G., 1955. Note on the bias of information estimates Information Theory. Psychology II-B ed. H Quastler (Glencoe, IL: Free Press). 95–100.

Montalvão, J., Silva, D.G., Attux, R. 2012. Simple entropy estimator for small datasets. Elec. Letters, 48, No. 17, 1059–1061.

Nemenman, I., Shafee, F., Bialek, W., 2002. Entropy and inference, revisited. T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, Adv. Neural Inf. Proc. Syst. 14, 1–9.

Nemenman, I., Bialek, W., de Ruyter van Steveninck, R., 2004. Entropy and information in neural spike trains: Progress on the sampling problem. Physical Review E, 1–7.

Nemenman, I., 2011. Coincidences and Estimation of Entropies of Random Variables with Large Cardinalities. Entropy. 13, 2013–2023.

Papoulis, A., Pillai, S. U., 2002. Probability, Random Variables, and Stochastic Processes. McGraw-Hill, 4th edition.

Rényi, A., 1960. On measures of information and entropy. in: Proceedings of the Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability, Berkeley, 547–561.

Jugurta Montalvão Jugurta Montalvão was born in Aracaju, Brazil, in 1968. He received the title of Electrical Engineer (1992) from the University of Campina Grande (UFPB II), Master in Electrical Engineering (1995) from the University of Campinas (UNICAMP) and Doctor in "Automatique et traitement du signal" (2000) from the University Paris-Sud XI. He joined the Department of Electrical Engineering of the Federal University of Sergipe (UFS) in 2005. His main research interests are: pattern recognition and signal processing.

Romis Attux Romis Attux was born in Goiânia, Brazil, in 1978. He received the titles of Electrical Engineer (1999), Master in Electrical Engineering (2001) and Doctor in Electrical Engineering (2005) from the University of Campinas (UNICAMP). Since 2007, he is an Assistant Professor (MS 3.2) at the same university. His main research interests are: unsupervised signal processing, computational intelligence, dynamical systems / chaos and brain-computer interfaces.

Daniel Guerreiro Daniel Guerreiro e Silva was born in Botucatu, Brazil, in 1983. He received the B.S. degree in computer engineering and the M.S. and Ph.D. degrees in electrical engineering, all from the University of Campinas (Unicamp), São Paulo, Brazil, in 2006, 2009, and 2013, respectively. Currently, he is a Professor at the Department of Electrical Engineering (ENE) of the University of Brasília (UnB). His main research interests are information theoretic learning, unsupervised signal processing and computational intelligence.