A measure of statistical complexity based on predictive information

Samer A. Abdallah and Mark D. Plumbley
Queen Mary University of London
(Dated: December 10, 2010)

We introduce an information theoretic measure of statistical structure, called 'binding information', for sets of random variables, and compare it with several previously proposed measures including excess entropy, Bialek *et al.*'s predictive information, and the multi-information. We derive some of the properties of the binding information, particularly in relation to the multi-information, and show that, for finite sets of binary random variables, the processes which maximises binding information are the 'parity' processes. Finally we discuss some of the implications this has for the use of the binding information as a measure of complexity.

arXiv:1012.1890v1 [math.ST] 8 Dec 2010

I. INTRODUCTION

The concepts of 'structure', 'pattern' and 'complexity' are relevant in many fields of inquiry: physics, biology, cognitive sciences, machine learning, the arts and so on; but are vague enough to resist being quantified in a single definitive manner. One approach, which we adopt here, is to attempt to characterise them in statistical terms, for *distributions* over configurations of some system, using the tools of information theory [1].

In this letter, we propose a measure of statistical structure based on the concept of *predictive information rate* (PIR) [2], which measures an aspect of temporal dependency not captured by previously proposed measures. We review a number of these earlier proposals and the PIR, and then define the *binding information* as the extensive counterpart of the PIR applicable to arbitrary countable sets of random variables. After describing some of its properties, we identify some finite discrete processes that maximise the binding information.

In the following, if X is a random process indexed by a set $\mathcal{A}$, and $\mathcal{B} \subseteq \mathcal{A}$, then $X_{\mathcal{B}}$ denotes the compound random variable (random 'vector') formed by taking X_α for each $\alpha \in \mathcal{B}$. The set of integers from M to N inclusive will be written $M..N$, and $\setminus$ will denote the set difference operator, so, for example, $X_{1..3\setminus\{2\}} \equiv (X_1, X_3)$.

II. BACKGROUND

Suppose that $(\ldots, X_{-1}, X_0, X_1, \ldots)$ is a bi-infinite stationary sequence of random variables, and that $\forall t \in \mathbb{Z}$, the random variable X_t takes values in a discrete set $\mathcal{X}$. Let μ be the associated shift-invariant probability measure. Stationarity implies that the probability distribution associated with any contiguous block of N variables $(X_{t+1}, \ldots, X_{t+N})$ is independent of t, and therefore we can define a shift-invariant block entropy function:

$$H(N) \triangleq H(X_1, \ldots, X_N) = \sum_{\mathbf{x} \in \mathcal{X}^N} -p_\mu^N(\mathbf{x}) \log p_\mu^N(\mathbf{x}), \quad (1)$$

where $p_\mu^N : \mathcal{X}^N \to [0,1]$ is the unique probability mass function for any N consecutive variables in the sequence, $p_\mu^N(\mathbf{x}) \triangleq \Pr(X_1 = x_1 \wedge \ldots \wedge X_N = x_N)$.

The *entropy rate* h_μ has two equivalent definitions in terms of the block entropy function [1, Ch. 4]:

$$h_\mu \triangleq \lim_{N\to\infty} \frac{H(N)}{N} = \lim_{N\to\infty} H(N) - H(N-1). \quad (2)$$

The block entropy function can also be used to express the mutual information between two contiguous segments of the sequence of length N and M respectively:

$$I(X_{-N..-1}; X_{0..M-1}) = H(N) + H(M) - H(N+M). \quad (3)$$

If we let both block lengths N and M tend to infinity, we obtain what has been called the *excess entropy* [3] or the *effective measure complexity* [4]. It is the amount of information about the infinite future that can be obtained, on average, by observing the infinite past:

$$E = \lim_{N\to\infty} 2H(N) - H(2N). \quad (4)$$

Bialek *et al.* [5] defined the *predictive information* $\mathcal{I}_{\text{pred}}(N)$ as the mutual information between a block of length N and the infinite future following it:

$$\mathcal{I}_{\text{pred}}(N) \triangleq \lim_{M\to\infty} H(N) + H(M) - H(N+M). \quad (5)$$

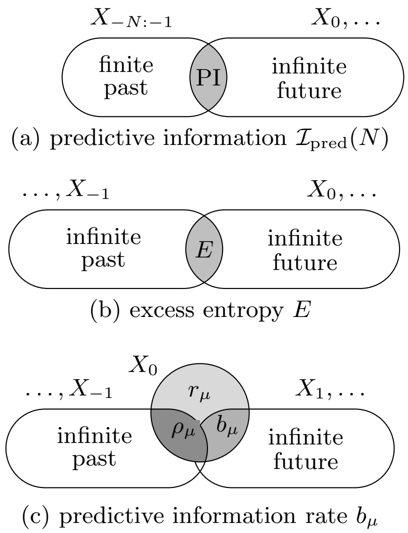
They showed that even if $\mathcal{I}_{\text{pred}}(N)$ diverges as N tends to infinity, the *manner* of its divergence reveals something about the learnability of the underlying random process. Bialek *et al.* also emphasised that $\mathcal{I}_{\text{pred}}(N)$ is the *sub-extensive* component of the entropy: if Nh_μ is the purely extensive (*i.e.*, linear in N) component of the entropy, then $\mathcal{I}_{\text{pred}}(N)$ is the difference between the block entropy $H(N)$ and its extensive component:

$$H(N) = Nh_\mu + \mathcal{I}_{\text{pred}}(N). \quad (6)$$

The *multi-information* [6] is defined for any collection of N random variables $(X_1, \ldots, X_N)$ as

$$I(X_{1..N}) \triangleq -H(X_{1..N}) + \sum_{i \in 1..N} H(X_i). \quad (7)$$

For $N = 2$, the multi-information reduces to the mutual information $I(X_1; X_2)$, while for $N > 2$, $I(X_{1:N})$ continues to be a measure of dependence, being zero if and only



(a) predictive information $\mathcal{I}_{\text{pred}}(N)$

(b) excess entropy E

(c) predictive information rate b_μ

FIG. 1. Venn diagram representation [1, Ch. 2] of several information measures for stationary random processes. Each circle or oval represents a random variable or sequence of random variables relative to time $t = 0$. Overlapped areas correspond to various mutual informations. In (c), the circle represents the 'present'. Its total area is $H(X_0) = H(1) = \rho_\mu + r_\mu + b_\mu$, where ρ_μ is the multi-information rate, r_μ is the residual entropy rate, and b_μ is the predictive information rate. The entropy rate is $h_\mu = r_\mu + b_\mu$.

if the variables are statistically independent. In the thermodynamic limit, the intensive *multi-information rate* (*cf.* Dubnov's *information rate* [7]) can be defined as

$$\rho_\mu \triangleq \lim_{N\to\infty} I(X_{1..N}) - I(X_{1..N-1}). \tag{8}$$

It can easily be shown that $\rho_\mu = \mathcal{I}_{\text{pred}}(1) = H(1) - h_\mu$. Erb and Ay [8] studied this quantity (they call it I) and showed that, in the present terminology,

$$I(X_{1..N}) + \mathcal{I}_{\text{pred}}(N) = N\rho_\mu. \tag{9}$$

Comparing this with (6), we see that $\mathcal{I}_{\text{pred}}(N)$ is also the sub-extensive component of the multi-information. Thus, all of the measures considered so far, being linearly dependent in various ways, are closely related.

Another class of measures, including Grassberger's *true measure complexity* [4] and Crutchfield *et al.*'s *statistical complexity* C_μ [9, 10], is based on the properties of stochastic automata that model the process under consideration. These have some interesting properties but are beyond the scope of this letter.

In [2], we introduced the *predictive information rate* (PIR), which is the average information in one observation about the infinite future given the infinite past. If $\overleftarrow{X}_t = (\ldots, X_{t-2}, X_{t-1})$ denotes the variables before time t, and $\overrightarrow{X}_t = (X_{t+1}, X_{t+2}, \ldots)$ denotes those after t, the PIR is defined as a conditional mutual information:

$$\overline{\underline{\mathcal{I}}}_t \triangleq I(X_t; \overrightarrow{X}_t | \overleftarrow{X}_t) = H(\overrightarrow{X}_t | \overleftarrow{X}_t) - H(\overrightarrow{X}_t | X_t, \overleftarrow{X}_t). \tag{10}$$

Equation (10) can be read as the average reduction in uncertainty about the future on learning X_t, given the past. Due to the symmetry of the mutual information, it can also be written as $\overline{\underline{\mathcal{I}}}_t = H(X_t | \overleftarrow{X}_t) - H(X_t | \overrightarrow{X}_t, \overleftarrow{X}_t)$. $H(X_t | \overleftarrow{X}_t)$ is the entropy rate h_μ, but $H(X_t | \overrightarrow{X}_t, \overleftarrow{X}_t)$ is a quantity that does not appear to be have been considered by other authors yet. It is the conditional entropy of one variable given *all* the others in the sequence, future as well as past. We call this the *residual entropy rate* r_μ, and define it as a limit:

$$r_\mu \triangleq \lim_{N\to\infty} H(X_{-N..N}) - H(X_{-N..-1}, X_{1..N}). \tag{11}$$

The second term, $H(X_{-N..-1}, X_{1..N})$, is the joint entropy of two non-adjacent blocks with a gap between them, and cannot be expressed as a function of block entropies alone. If we let b_μ denote the shift-invariant PIR, then $b_\mu = h_\mu - r_\mu$ (see Fig. II).

Many of the measures reviewed above were intended as measures of 'complexity', a quality that is somewhat open to interpretation [11, 12]. It is generally agreed, however, that complexity should be low for systems that are deterministic or easy to compute or predict—'ordered'—and low for systems that a completely random and unpredictable—'disordered'. The PIR satisfies these conditions without being 'over-universal' in the sense of Crutchfield *et al.* [12, 13]: it is not simply a function of entropy or entropy rate that fails to distinguish between the different strengths of temporal dependency that can be exhibited by systems at a given level of entropy. In our analysis of Markov chains [2], we found that processes which maximise the PIR do not maximise the multi-information rate ρ_μ (or the excess entropy, which is the same in this case), but do have a certain kind of partial predictability that requires the observer continually to pay attention to the most recent observations in order to make optimal predictions. And so, while Crutchfield *et al.* make a compelling case for the excess entropy E and their statistical complexity C_μ as measures of complexity, there is still room to suggest that the PIR captures a different and non trivial aspect of temporal dependency structure not previously examined.

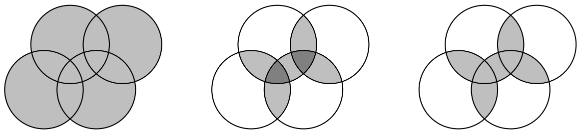
III. BINDING INFORMATION

If the PIR rate is accumulated over successive time steps, a quantity which we call the *binding information* is obtained. To proceed, we first reformulate the infinite sequence PIR (10) so that it becomes applicable to a *finite* sequence of random variables $(X_1, \ldots, X_N)$:

$$\overline{\underline{\mathcal{I}}}_t(X_{1..N}) = I(X_t; X_{(t+1)..N} | X_{1..(t-1)}), \tag{12}$$

Note that this is no longer shift-invariant and may depend on t. The binding information, then, is the sum

$$B(X_{1..N}) = \sum_{t\in 1..N} \overline{\underline{\mathcal{I}}}_t(X_{1..N}). \tag{13}$$



(a) $H(X_{1..4})$ (b) $I(X_{1..4})$ (c) $B(X_{1..4})$

FIG. 2. Illustration of binding information as compared with multi-information for a set of four random variables. In each case, the quantity is represented by the total amount of black ink, as it were, in the shaded parts of the diagram. Whereas the multi-information counts the multiply-overlapped areas multiple times, the binding information counts each overlapped areas just once.

Expanding this in terms of entropies and conditional entropies and cancelling terms yields

$$B(X_{1..N}) = H(X_{1..N}) - \sum_{t\in 1..N} H(X_t|X_{1..N\setminus\{t\}}). \quad (14)$$

Like the multi-information, it measures dependencies between random variables, but in a different way (see fig. III). Though the binding information was derived by accumulating the PIR sequentially, the result is permutation invariant, suggesting that the concept might be applicable to arbitrary sets of random variables regardless of their topology. Accordingly, we define the binding information as follows:

Definition 1. *If $\{X_\alpha|\alpha\in\mathcal{A}\}$ is set of random variables indexed by a countable set $\mathcal{A}$, the binding information is*

$$B(X_{\mathcal{A}}) \triangleq H(X_{\mathcal{A}}) - \sum_{\alpha\in\mathcal{A}} H(X_\alpha|X_{\mathcal{A}\setminus\{\alpha\}}). \quad (15)$$

Since the binding information can be expressed as a sum of (conditional) mutual informations between sets of random variables (13), it is (a) non-negative and (b) invariant to invertible pointwise transformations of the variables; that is, if $Y_{\mathcal{A}}$ is a set of random variables such that, $\forall\alpha\in\mathcal{A}$, $Y_\alpha = f_\alpha(X_\alpha)$ for some invertible functions f_α, then $B(Y_{\mathcal{A}}) = B(X_{\mathcal{A}})$.

The binding information is zero for sets of independent random variables—the case of complete 'disorder'—and zero when all variables have zero entropy, taking known values and representing a certain kind of 'order'. However, it is also possible to obtain low binding information for *random* systems which are nonetheless very ordered in a certain way. If each variable X_α is some function of $X_{\alpha'}$ for all $\alpha'\neq\alpha$, then the state of the entire system can be read off from any one of its component variables. In this case, it is easy to show that $B(X_{\mathcal{A}}) = H(X_{\mathcal{A}}) = H(X_\alpha)$ for any $\alpha\in\mathcal{A}$, which, as we will see, is relatively low compared with what is possible as soon as N becomes appreciably large. Thus, binding information is low for both highly 'ordered' and highly 'disordered' systems, but in this case, 'highly ordered' does *not* simply mean deterministic or known *a priori*: it means the whole is predictable from the smallest of its parts.

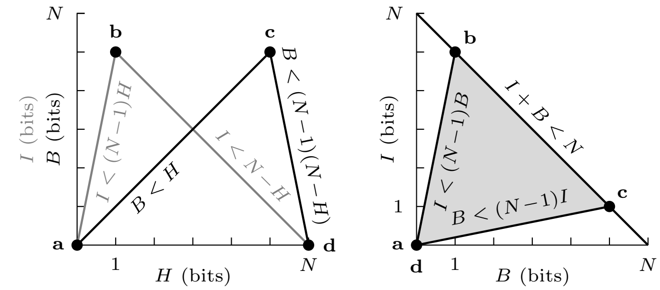


FIG. 3. Constraints on multi-information $I(X_{1..N})$ and binding information $B(X_{1..N})$ for a system of $N=6$ binary random variables. The labelled points represent identifiable distributions over the 2^N states that this system can occupy: (a) *known state*, the system is deterministically in one configuration; (b) *giant bit*, one of the $P^6_{\mathcal{B}}$ processes; (c) *parity*, the parity processes $P^6_{2,0}$ or $P^6_{2,1}$; (d) *independent*, the system of independent unbiased random bits.

IV. BOUNDS ON BINDING AND MULTI-INFORMATION

In this section we confine our attention to sets of discrete random variables taking values in a common alphabet containing K symbols. In this case, it is quite straightforward to derive upper bounds, as functions of the joint entropy, on both the multi-information and the binding information, and also upper bounds on multi-information and binding information as functions of each other. In [14], we prove the following results:

Theorem 1. *If $\{X_\alpha|\alpha\in\mathcal{A}\}$ is a set of $N=|\mathcal{A}|$ random variables all taking values in a discrete set of cardinality K, then the following constraints all hold:*

$$I(X_{\mathcal{A}}) \le N\log K - H(X_{\mathcal{A}}) \quad (16)$$
$$I(X_{\mathcal{A}}) \le (N-1)H(X_{\mathcal{A}}) \quad (17)$$
$$B(X_{\mathcal{A}}) \le H(X_{\mathcal{A}}) \quad (18)$$
$$B(X_{\mathcal{A}}) \le (N-1)(N\log K - H(X_{\mathcal{A}})). \quad (19)$$

Also, $B(X_{\mathcal{A}})$ and $I(X_{\mathcal{A}})$ are mutually constrained:

$$I(X_{\mathcal{A}}) + B(X_{\mathcal{A}}) \le N\log K. \quad (20)$$

These bounds restrict $I(X_{\mathcal{A}})$ and $B(X_{\mathcal{A}})$ to two triangular regions of the plane when plotted against the joint entropy $H(X_{\mathcal{A}})$ and are illustrated for $N=6, K=2$ in fig. IV. Two more linear bounds were suggested by empirical computations of binding information and multi-information:

$$I(X_{\mathcal{A}}) \le (N-1)B(X_{\mathcal{A}}) \quad (21)$$
$$\text{and}\quad B(X_{\mathcal{A}}) \le (N-1)I(X_{\mathcal{A}}). \quad (22)$$

We have not found a general proof of these inequalities for all N, but we have constructed a numerical algorithm [14] that is able to find proofs for given values of N up to 37, at which point insufficient numerical precision becomes the limiting factor.

V. MAXIMISING BINDING INFORMATION

Is the absolute maximum of $B(X_{1..N}) = (N-1)\log K$ implied by Theorem 1 is attainable, and by what kinds of processes? In [14] we prove the following:

Theorem 2. *If $\{X_1, \ldots, X_N\}$ is a set of discrete random variables each taking values in $0..(K-1)$, then $B(X_{1..N})$ is maximised at $(N-1)\log_2 K$ bits by the K 'modulo-K processes' $P^N_{K,m}$ for $m \in 0..(K-1)$, under which the probability of a configuration $\mathbf{x} \in (0..K-1)^N$ is*

$$P^N_{K,m}(\mathbf{x}) = \begin{cases} K^{1-N} & \textit{if } \left(\sum_{i=1}^N x_i\right) \bmod K = m, \\ 0 & \textit{otherwise.} \end{cases} \tag{23}$$

When $K = 2$ (binary random variables) the maximal binding information of $N-1$ bits is reached by the two 'parity' processes: $P^N_{2,0}$ is the 'even' process, which distributes uniform probability over all configurations with even parity; $P^N_{2,0}$ is the 'odd' process, which distributes uniform probabilities over the complementary set. The multi-information of the parity processes is 1 bit. By contrast, the binary processes which maximise the multi-information at $N-1$ bits are the 'giant bit' processes: the indices $1..N$ are partitioned into two sets $\mathcal{B}$ and its complement $\overline{\mathcal{B}} = 1..N \setminus \mathcal{B}$, and probabilities assigned to configurations $\mathbf{x} \in \{0,1\}^N$ as follows:

$$P^N_{\mathcal{B}}(\mathbf{x}) = \begin{cases} \frac{1}{2} & : \text{if } \forall i \in 1..N \,.\, x_i = \mathbb{I}(i \in \mathcal{B}), \\ \frac{1}{2} & : \text{if } \forall i \in 1..N \,.\, x_i = \mathbb{I}(i \in \overline{\mathcal{B}}), \\ 0 & : \text{otherwise,} \end{cases} \tag{24}$$

where $\mathbb{I}(\cdot)$ is 1 if its argument is true and 0 otherwise. The binding information of these processes is 1 bit. Thus we see that the processes which maximise the binding information and the multi-information are quite different in character.

VI. DISCUSSION AND CONCLUSIONS

As noted in § I, Bialek *et al.* argue that the predictive information $\mathcal{I}_{\text{pred}}(N)$, being the sub-extensive component of the entropy, is the unique measure of complexity that satisfies certain reasonable desiderata, including transformation invariance for continuous-valued variables [5, §5.3]. While lack of space precludes a full discussion, we note that transformation invariance does *not*, as Bialek *et al.* state [5, p. 2450], demand sub-extensivity: binding information *is* transformation invariant, since it is a sum of conditional mutual informations, and yet it *can* have an extensive component, since its intensive counterpart, the PIR, can have a well-defined value, *e.g.*, in stationary Markov chains [2].

Measures of statistical dependency are discussed by Studenỳ and Vejnarovà, [6, §4], who formulate a 'level-specific' measure that captures the dependency visible when *fixed size* subsets of variables are examined in isolation. Studenỳ and Vejnarovà [6, p. 277] use the parity process as an example of a random process in which the dependence is only visible at the highest level, that is, amongst all N variables; if fewer than N variables are examined, they appear to be independent. They note that such processes were called 'pseudo-independent' by Xiang *et al.* [15], who concluded that standard algorithms for Bayesian network construction fail when applied to them. It is intriguing, then, that these are singled out as 'most complex' according to the binding information criterion.

To summarise, we have introduced binding information as a measure of statistical structure that can be applied to any countable set of random variables regardless of any topological organisation of the variables. Binding information is maximised in finite discrete valued systems by the 'modulo process'. Further results on binding information, and investigations of binding information in some specific random processes are presented in [14].

---

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (John Wiley and Sons, New York, 1991).

[2] S. A. Abdallah and M. D. Plumbley, Connection Science **21**, 89 (2009).

[3] J. Crutchfield and N. Packard, Physica D: Nonlinear Phenomena **7**, 201 (1983).

[4] P. Grassberger, International Journal of Theoretical Physics **25**, 907 (1986).

[5] W. Bialek, I. Nemenman, and N. Tishby, Neural Computation **13**, 2409 (2001).

[6] M. Studenỳ and J. Vejnarovà, in *Learning in Graphical Models*, edited by M. I. Jordan (MIT Press, 1998) pp. 261–297.

[7] S. Dubnov, Signal Processing Letters, IEEE **11**, 698 (2004).

[8] I. Erb and N. Ay, Journal of Statistical Physics **115**, 949 (2004).

[9] J. P. Crutchfield and K. Young, Physical Review Letters **63**, 105 (1989).

[10] J. P. Crutchfield and D. P. Feldman, Physical Review E **55**, 1239R (1997).

[11] C. H. Bennett, in *Complexity, Entropy, and the Physics of Information*, edited by W. H. Zurek (Addison-Wesley, 1990) pp. 137–148.

[12] D. P. Feldman and J. P. Crutchfield, Physics Letters A **238**, 244 (1998).

[13] J. P. Crutchfield, D. P. Feldman, and C. R. Shalizi, Physical Review E **62**, 2996 (2000).

[14] S. A. Abdallah and M. D. Plumbley, *Predictive Information, Multi-information and Binding Information*, Tech. Rep. C4DM-TR-10-10 (Queen Mary University of London, 2010).

[15] Y. Xiang, S. Wong, and N. Cercone, in *Proc. 12th Conf. on Uncertainty in Artificial Intelligence* (1996) pp. 564–571.