SYMBOLIC DYNAMIC ANALYSIS THROUGH COINCIDENCE DETECTION – AN OVERSIMPLIFIED METHOD

Jugurta Montalvão*, Romis Attux†, Daniel Silva†

*Universidade Federal de Sergipe (UFS)
São Cristóvão, Sergipe, Brazil

†University of Campinas (UNICAMP), São Paulo, Brazil

Emails: jmontalvao@ufs.br, attux@dca.fee.unicamp.br, danielgs@dca.fee.unicamp.br

Abstract— A very pragmatic approach for measuring memory (or inertia) in dynamic sequences of symbols (symbol dynamics) is proposed, where only intervals between coincidences of symbols along the sequence are taken into account in the process of estimating the Auto Mutual Information. The proposed method is studied using sequences of symbols obtained from two Markov sources, with two and nine states respectively. Results are compared to expected theoretical values of mutual information, as well as to histogram-based estimations with Miller's entropy bias compensation.

Keywords— Auto Mutual Information, Coincidence detection, Dynamic memory.

Stochastic processes (or random functions) are ubiquitous tools in Science, frequently employed as powerful mathematical models when random behaviours – through time or space – are observed in numerically represented signals. As illustrated in Figure 1, both natural (e.g. biological data) and human-made (e.g. industrial plants) systems can be regarded as sources of random signals, and these signals are typically understood as instances of stochastic processes (Papoulis and Pillai, 2002).

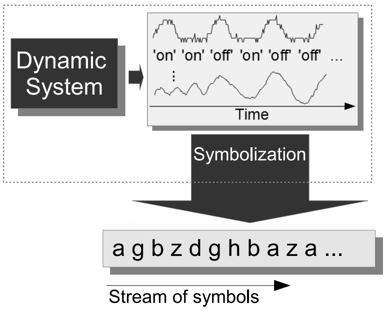
Dynamic System
'on' 'on' 'off' 'on' 'off' 'off' ...
Time
Symbolization
a g b z d g h b a z a ...
Stream of symbols

Figure 1: Symbolization of signals from dynamic sources.

On the other hand, observed dynamics in areas such as social science, public health, zoology, education and marketing, are commonly categorical (Agresti, 2007; Daw et al., 2003). For instance, just to give a brief flavor of what this means, we may consider dynamic cycles of "euphoria" and, we may consider dynamic cycles of "euphoria" and "depression" – possibly including subcategories of these mental states –, in psychology, or sequences of base pairs belonging to a certain DNA chain, in biology.

Moreover, even in highly quantitative fields such as engineering sciences and industrial quality control, the use of symbols instead of numbers can simplify dynamic analysis. This reasoning gives support to the categoric data analysis, or symbolic analysis, already used for many years in astrophysics, health, mechanics, artificial intelligence, control, telecommunications and data mining (Daw et al., 2003). In most such cases, the mapping of numeric signals into symbols is called "Symbolization", which clearly includes vector-quantization. For the interested reader, we suggest the work by Sant'Anna and Wickström (2011) (Sant'Anna and Wickström, 2011), where some symbolization methods are cross-compared.

In this work, we do not address symbolization issues. Instead, we assume that whatever be the underlying dynamic, a suitable symbolization was applied so that all observations are streams of symbols, as illustrated in Figure 1.

Unfortunately, unlike in stochastic processes analysis, pivotal tools such as auto-correlation and power spectrum are not defined in Symbolic Analysis (SA). By contrast, Auto-Mutual Information (AMI) is a quite straightforward option as a first step to SA. More specifically, given an infinitely long sequence of symbols, $x_{-\infty}^{+\infty}$, $x_i \in S = \{s_1, s_2, \ldots, s_K\}$, where K is the cardinality of the sampling set, for any Integer τ, and a randomly chosen position n, the Auto Mutual Information is given, in *bits* of information, by the following Entropy balance (Cover and Thomas, 1991):

$$I(\tau, n) = H(X_n) + H(X_{n+\tau}) - H(X_n, X_{n+\tau}) \quad (1)$$

where $H(X_n) = -\sum_{i=1}^{K} P(X_n = s_i) \log_2 P(X_n = s_i)$ and $H(X_n, X_{n+\tau}) = \sum_{i=1}^{K} \sum_{j=1}^{K} P(X_n = s_i, X_{n+\tau} = s_j) \log_2 P(X_n = s_i, X_{n+\tau} = s_j)$.

By considering the stationary case, where probabilities $P(X_n = s_i)$ and $P(X_n = s_i, X_{n+\tau} = s_j)$ do not depend on n, it turns out that $I(\tau)$ does not depend on n too, and can be conveniently estimated by the replacement of probability entries, in Eq. 1, with relative frequencies of symbols. In other words, to estimate entropy, a natural approach is to take as many samples as possible to build *histograms* and then to use these histograms as probabilities entries in Shannon's formula. These approaches are known as plug-in methods (Beirlant et al., 1997).

As a matter of fact, since practical instances of symbol streams are finite, relative frequencies are random variables whose logarithm transformation, in Eq. 1, induces systematic biases in entropy estimation. For not too small streams of symbols, bias compensation such as that proposed by G. Miller (Miller, 1955) can improve estimation quality. In this paper, we systematically apply Miller's compensation in all AMI analysis with plug-in methods. For instance, in Figure 3, we compare the theoretical values of Normalized AMI (NAMI),

$$NAMI(\tau) = \frac{H(X_n) + H(X_{n+\tau}) - H(X_n, X_{n+\tau})}{H(X_n)} \quad (2)$$

as a function of τ, to estimated values trough two estimation methods, including the usual plug-in approach with Miller's bias compensation (Miller, 1955). In this illustration, we use the very simple two-state Markov source represented in Figure 2. There, we can note that the NAMI approaches zero for $\tau > 25$, what indicates that a symbolic event that occurs at a certain time n has almost no influence over events after more than 25 discrete intervals (supposing regular interval between samples). Furthermore, from another point of view, we can infer that Markov's transition probability matrix (Cover and Thomas, 1991) in q steps, P^q, is almost the same for all $q \geq 25$. We can refer to this decrease in event influence through time as a 'memory' or an 'inertia' of the underlying dynamics. By itself, the inference of memory/inertia in Symbolic Dynamics justifies our interest in AMI estimators (consider, for instance, the impact of properly measuring the extent of influence of a given event through time in Economics of Social Dynamics).

Unfortunately, though very easy to be understood and coded in any programming language, the plug-in approach to be properly applied, tacitly demands stochastic sources taking symbols from sets with moderate cardinality, K, and long enough samples of symbol streams, x_1^L, thus satisfying the trade-off $L >> K^2$. Otherwise, the estimation of the joint probability $P(X_n = s_i, X_{n+\tau} = s_j)$ would not be reliable enough.

In 2012, we proposed a very simple method (Montalvão et al., 2012) to estimate entropy whose main advantage, besides its simplicity of use, is that it can estimate entropy even with $L < K$. For the reader's convenience, we revisit here the essential explanations from (Montalvão et al., 2012) in Section 1. Then, in Section 2, we propose a first (non-simplified) version of a Normalized AMI estimator based on Coincidence Detection, which is eventually (over)simplified to become the method proposed in Section 3. Illustrative results are presented in Section 4.

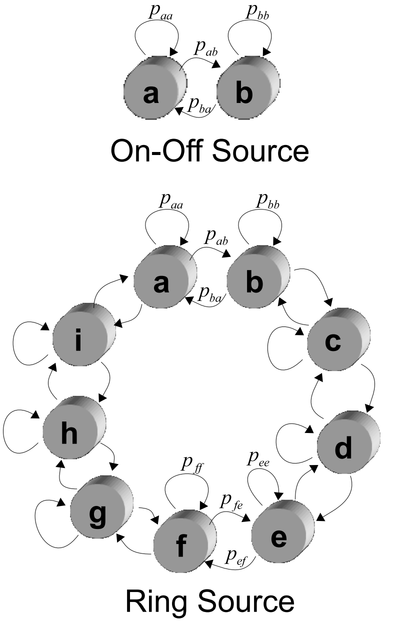

Figure 2: Markovian symbol sources used in this paper.

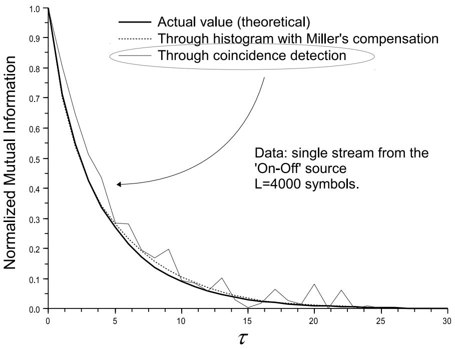

Figure 3: Auto Mutual Information of Source 'On-Off', with $p_{ab} = p_{ba} = 0.05$.

1 From the 'Birthday Problem' to Entropy Estimation

Motivated by the practical limitations of plug-in methods, we formulated the following question: 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 gather coincidences of each symbol in a stream of symbols. Though, using K coincidence detectors can be costly.

By contrast, a method of entropy estimation through coincidences was proposed in 1985 by Ma (Ma, 1985, Ch. 25) as a 'method (...) in the stage of development', to be used in Statistical Mechanics. The author also discusses an interesting link between information theory 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 common perception by the time his book was written. Nonetheless, in some hard problems involving blocks of symbols, such as Multiple Input / Multiple Output digital systems, even small sets of symbols may lead to problems of entropy estimation with a huge number of states, not to mention Symbolic Dynamics to mention Symbolic Dynamic analysis in new challenging problems such as Data Mining in databases with a large number of attributes. In other words, nowadays, Ma-like methods can be appropriate for a myriad of very relevant problems beside statistical mechanics.

Thus, 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 of symbols, this unconstrained coincidence detection is closely related to the classical 'Birthday Problem', presented in textbooks of probability (Papoulis and Pillai, 2002). By generalizing this problem, let K be the number of equiprobable symbols, if they are drawn from this source, the probability of repeating one or more symbols by the n-th sample is given by:

$$F(n;K) = 1 - \frac{K(K-1)(K-2)\dots(K-n+1)}{K^n} \tag{3}$$

were $n = 1, 2, \dots, K$ and K plays the role of a parameter for this accumulated probability distribution. Therefore, the probability of a first coincidence precisely at the n-th sample, for $1 < n \leq K$, is given by $f(n;K) = F(n;K) - F(n-1;K)$. We can further estimate the average number of samples drawn from the source until a first coincidence occurs as:

$$D(K) = \sum_{n=0}^{K} n f(n;K) \tag{4}$$

which clearly depends on K. For instance, in the Birthday Problem itself, considering $K = 365$ days, on average, we shall expect one birthday coincidence roughly every 24 independently consulted subjects. Figure 4 graphically presents D as a function of K, from $K = 2$ to $K = 2000$. By inverting the axis in Figure 4, we can see a striking quadratic functional dependence of K on D. Indeed, by adjusting the polynomial model:

$$\hat{K}(D) = aD^2 + bD + c \tag{5}$$

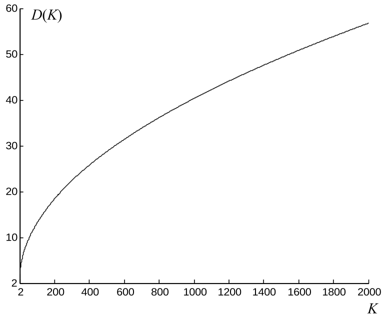

Figure 4: Averaged number of symbols, D, drawn from a source of K equiprobable symbols until a first coincidence occurs.

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 $\hat{K}$ of about 10^{-6}, inside the interval $D(1) = 2$ to $D(2000) \approx 56.7$. This polynomial approximation is a key aspect of our simplified method.

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 non-equiprobable sources of symbols. Clearly, we still may empirically estimate $\hat{D}$ by sequentially observing symbols and averaging the number of symbols between coincidences. 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 until first coincidence. Therefore, we should expect $\hat{K} \approx 2^H$.

As a result, our very pragmatic method for entropy estimation was summarized in three steps:

1 Estimate D by sequential observation of symbols, as illustrated in Figure 5, 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 $\hat{H} = \log_2(\hat{K})$.

To illustrate this method at work, we chose the emblematic source of 365 equiprobable symbols from the Birthday Problem, and measured its bias in several scenarios. In this case, the known source entropy is $H = \log_2(365) = 8.5118$ bits, for the second column in Table 1, where this (equiprobable) source was simulated and $\hat{D}$ was obtained through the observation of N sequential symbols (with at least one coincidence). Then, we

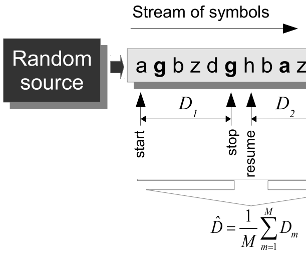
Stream of symbols
Random source
a g b z d g h b a z a ...
D_1
D_2
...
start
stop
resume
stop
$\hat{D} = \frac{1}{M}\sum_{m=1}^{M} D_m$

Figure 5: Incremental estimation of the averaged number of symbols until coincidence detection, where $D_1 = 6$ and $D_2 = 5$.

applied the proposed method and calculated the bias $\hat{H} - H$ for 10^4 independent trials. Similarly, in the third column, we present the average bias for sources whose probability distributions were randomly generated (thus yielding $H \leq \log_2(365)$).

Table 1: Average estimation of relative biases $\left(\frac{\hat{H}-H}{H}\right)$, for memoryless sources of $K = 365$ symbols, after N sequential observations.

N	Average relative bias (uniform distribution)	Average relative bias (random distributions)
50	-0.042	-0.046
70	-0.022	-0.033
90	-0.014	-0.028
100	-0.013	-0.028
200	-0.005	-0.021
500	-0.002	-0.018
1000	-0.001	-0.015

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 sources of symbols, the value of $\hat{D}$ can be iteratively improved, even when K is not known.

2 Proposed Method for Normalized AMI Estimation

Given a finite sequence of symbols, x_1^L, for every integer τ ($|\tau| << L$), a new sequence $y_1^{L-\tau}$ is formed by symbol concatenation, so that $y_n = x_n \oplus x_{n+|\tau|}$, where $\oplus$ stands for symbol concatenation operator. Consequently, $y_n \in S \times S$. This step is illustrated in Figure 6.

The proposed method is based on the averaged number of symbols until a first coincidence occurs in both x and y, but a very important aspect of our method is that it assumes that symbols are independently drawn. Evidently, this is not true for sequences where estimation of AMI is worth it (exactly because of its 'memory'). Therefore, we include a step corresponding to the random permutation of both x and y, thus yielding two 'white' sequences, x, y, as illustrated in Figure 7. Accordingly, the proposed method for AMI estimation can be implemented through the following algorithm:

1 Given a sequence x_1^L, for every τ ($|\tau| << L$), obtain a new sequence $y_1^{L-\tau}$, formed by symbol concatenation ($y_n = x_n \oplus x_{n+|\tau|}$) (see Figure 6).

2 Randomly permute x_1^L and $y_1^{L-\tau}$ to obtain x_1^L and $\mathsf{y}_1^{L-\tau}$, respectively (see Figure 7).

3 Estimate D_x and D_y, respectively, by sequential observation of coincidences in x_1^L and $\mathsf{y}_1^{L-\tau}$, respectively, as illustrated in Figure 5, thus obtaining $\hat{D}_x$ and $\hat{D}_y$.

4 Compute $\hat{K}_x(\hat{D}) = a\hat{D}_x^2 + b\hat{D}_x + c$ and $\hat{K}_y(\hat{D}) = a\hat{D}_y^2 + b\hat{D}_y + c$, with $a = 0.6366$, $b = -0.8493$ and $c = 0.1272$.

5 Estimate entropies, in bits, as $\hat{H}_x = \log_2(\hat{K}_x)$ and $\hat{H}_y = \log_2(\hat{K}_y)$.

6 Estimate $NAMI(\tau) = (2\hat{H}_x - \hat{H}_y)/\hat{H}_x$.

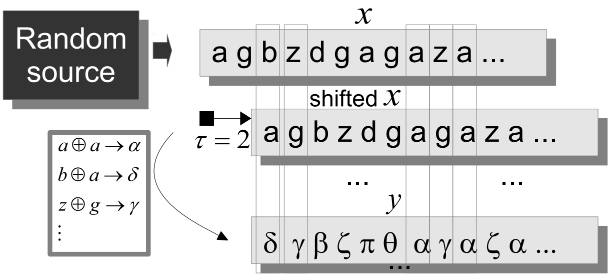

Figure 6: Synthesis of sequence y for $\tau = 2$.

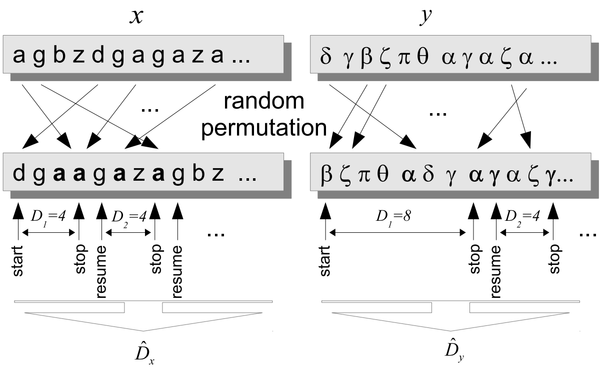

Figure 7: Random permutation of symbol sequences followed by estimation of average interval between coincidence detections.

To illustrate the performance of this proposed method, we recall that the rugged line in Figure 3 corresponds to the NAMI estimated for a single sequence of 4000 symbols from the Markovian source 'On-Off'. The rugged effect comes as a consequence of that most symbols are 'discarded' in a single run of the method. In other words, unlike histograms, where every symbol observed is taken into account, our method discards symbols between coincidences. However, this can be easily compensated for, as a single sequence of L symbols allows up to $L!$ random permutations (step 2 of the method); therefore, just by moving back to this step 2 a few times and averaging obtained results allows a smoothing of the NAMI curve, as illustrated, in Figure 8, for the very same sequence of 4000 symbols from the source 'On-Off'.

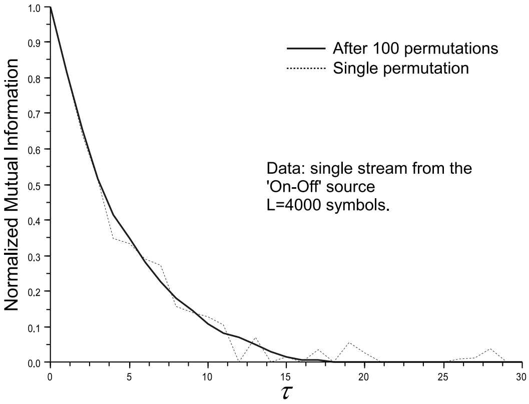

Figure 8: Smoothing effect of multiple permutations – Source 'On-Off', with $p_{ab} = p_{ba} = 0.05$.

3 Oversimplified version of the proposed Method for AMI Estimation

Estimation of $NAMI(\tau) = (2\hat{H}_x - \hat{H}_y)/\hat{H}_x$ (or $NAMI(\tau) = 2 - \hat{H}_y/\hat{H}_x$), in our method, entirely relies upon average observation of intervals between symbol coincidences, $\hat{D}_x$ and $\hat{D}_y$, in two sequences. By keeping simplicity as leitmotiv, we can push even further the polynomial approximation of $\hat{K}(D)$, in Equation 5. Indeed, since $\approx (0.6366D^2 - 0.8493D + 0.1272) \approx (D-1)^{\alpha}$, with $\alpha \approx 1.85$ for values shown in Figure 4, we can oversimplify the estimation by approximating

$$NAMI(\tau) \approx 2 - \frac{\log_2(\hat{D}_y - 1)}{\log_2(\hat{D}_x - 1)} \qquad (6)$$

.

4 Results

To highlight the effect of the quadratic increase (K^2) of cardinality of the set from which sequence y takes symbols, besides the 'on-off' source with 2 states, we also use a ring-like Markov chain with 9 states ('Ring' source), both already illustrated in Figure 2.

As noticed in Figure 9, unlike the histogram-based approach, our method does not adopt bias compensation, and, consequently, is outperformed by plug-in methods.

our method is not bias compensated, consequently it is outperformed by plug-in methods.

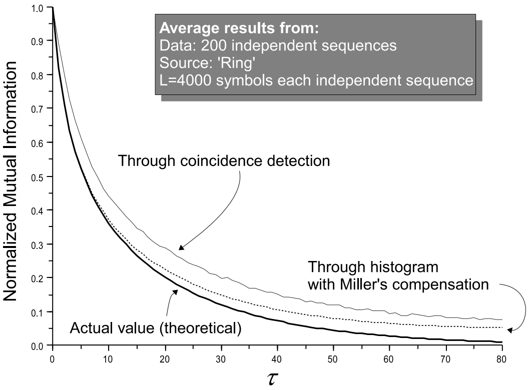

Figure 9: Auto Mutual Information of Source 'Ring', with $p_{ij} = p_{ji} = 0.05, i \neq j$, 4000 symbols in each sequence. Presented results correspond to the average AMI after 200 independent trial.

By contrast, in Figure 10, we observe that, under data shortage (i.e. $L = 100$ whereas the cardinality of the set from which y takes symbols is $K^2 = 9^2$) the coincidence-based approach is less disturbed than the plug-in strategy.

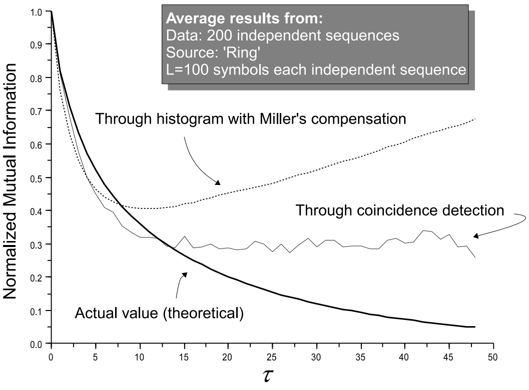

Figure 10: Auto Mutual Information of Source 'Ring', with $p_{ij} = p_{ji} = 0.05, i \neq j$ and only 100 symbols in each sequence. Presented results correspond to the average AMI after 200 independent trial.

Finally, to illustrate the performance of the oversimplified estimator, in Figure 11, we repeat experiments with the 'Ring' source. In spite of the exposed simplifications, we can see that, as for the coincidence-detection based method, that differences in the results are barely perceptible, by comparison to results in Figure 9, also presenting a much less degraded performance under data shortage than the plug-in approach (the stronger performance degradation of the histogram-based method is indicated with an arrow in Figure 11).

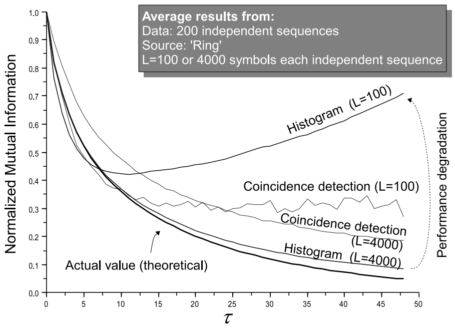

Figure 11: Auto Mutual Information of Source 'Ring', with $p_{ij} = p_{ji} = 0.05, i \neq j$. Presented results compare the performances of the histogram based approach to the Oversimplified Method based on coincidence detection, highlighting the strong degradation of the former under data shortage.

5 Conclusions

In this brief work, we extended the application of an entropy estimator based on coincidence to Auto Mutual Information analysis. This proposal has two main attractive aspects: (a) it is very simple to use, for it is based on the observation of intervals between coincidences, and (b) it is robust to data shortage, because entropy can be estimated, through coincidence detection, even when the number of available observations is less than the symbol set cardinality. This robustness was illustrated through experiments with simple Markov sources.

Moreover, by keeping simplicity as leitmotiv, we even oversimplified the estimator to an expression where a direct comparison of intervals between coincidences is explicit. We found this result particularly interesting, from an intuitive point of view, because we now can think of the Auto Mutual Information (hence of the memory/inertia of a given dynamic translated into symbols) as a simple comparison between (time or space) intervals in a logarithmic scale. Note also that the comparison of intensities in logarithmic scales is a quite human-like approach, being part of human reasoning, which naturally raises the possibility of existence of links between human perception of memory/inertia and the oversimplified version of our NAMI estimator.

On the other rand, our results also show that, in its current stage of development, our method is not bias-free. Indeed, we are now working on two main subjects: i) bias compensation and ii) extension toward detection of coincidences defined in continuous spaces.

Acknowledgements

This work was partially supported by CNPq (303707/2012-7 and 302890/2012-2) and CAPES.

References

Agresti, A. (2012). An Introduction to Categorical Data Analysis, Wiley, New Jersey.

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

Cover, T. M. and Thomas, J. A. (1991). Elements of Information Theory, Wiley, New Jersey.

Daw, C. S. and C. E. A. Finney and E. R. Tracy. (2003). A review of symbolic analysis of experimental data., *Review of Scientific Instruments.* **74:916–930**.

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

Miller, G.,(1985). Note on the bias of information estimates Information Theory., *Psychology II-B ed. H Quastler (Glencoe, IL: Free Press).* **95–100**.

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

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

Sant'Anna, A. and Wickström, N. 2011. (2012). Symbolization of time series : an evaluation of SAX, persist, and ACA., *Proceedings of the 4th International Congress on Image and Signal Processing (CISP 2011), 15-17 October 2011, Shanghai, China,* **2223–2228**.