Predictive information in Gaussian processes with application to music analysis

Samer Abdallah[1] and Mark Plumbley[2]

[1] University College London
[2] Queen Mary University of London

Abstract. We describe an information-theoretic approach to the analysis of sequential data, which emphasises the predictive aspects of perception, and the dynamic process of forming and modifying expectations about an unfolding stream of data, characterising these using a set of process information measures. After reviewing the theoretical foundations and the definition of the predictive information rate, we describe how this can be computed for Gaussian processes, including how the approach can be adpated to non-stationary processes, using an online Bayesian spectral estimation method to compute the Bayesian surprise. We finish with a sample analysis of a recording of Steve Reich's *Drumming*.

1 Introduction

The concept of predictive information in a random process has developed over a number of years, with many contributions to be found in the physics and machine learning literature. For example, the *excess entropy* [1] is the mutual information between the semi-infinite past and future of a random process. Addressing the observation that some processes with long-range dependencies have infinite excess entropy [2], Bialek *et al* [3] introduced the *predictive information* as the mutual information between a *finite* segment of a process and the infinite future following it, and studied its behaviour, especially in relation to learning in statistical models. In previous work [4], we defined the *predictive information rate* (PIR) of a random process as the average information in one observation about future observations yet to be made *given* the observations made so far; thus, it quantifies the *new* information in observations made sequentially. The PIR captures a dimension of temporal structure that is not accounted for by previously proposed measures. In this paper, we show how various process information measures including the PIR are defined for discrete-time Gaussian processes, and apply this to the analysis of musical audio using an adaptive nonstationary Gaussian process model.

2 Information measures for stationary random processes

For an infinite stationary discrete-time random process $(X_t)_{t\in\mathbb{Z}}$, the predictive information rate (PIR), as defined in [4], is global measure of temporal structure that characterises the process, or statistical ensemble, as a whole, rather than

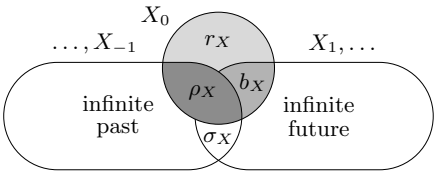

Fig. 1. I-diagram representation of several information measures for stationary random processes. Each circle or oval represents one or more random variables. The circle represents the 'present'. Its total area is $H(X_0) = \rho_X + r_X + b_X$, where ρ_X is the multi-information rate, r_X is the erasure entropy rate, and b_X is the predictive information rate. The entropy rate is $h_X = r_X + b_X$. The excess entropy is $E_X = \rho_X + \sigma_X$.

for particular realisations of the process, in the same way that the entropy rate characterises its overall randomness. In previous work [5] we examined several process information measures and their interrelationships, as well as generalisation of these for arbitrary countable sets of random variables. Following the conventions established there, we let $\overleftarrow{X}_t = (\ldots, X_{t-2}, X_{t-1})$ denote the variables before time t, and $\overrightarrow{X}_t = (X_{t+1}, X_{t+2}, \ldots)$ denote those after t. The predictive information rate b_X of the process X is defined as the conditional mutual information

$$b_X = I(X_t; \overrightarrow{X}_t | \overleftarrow{X}_t) = H(\overrightarrow{X}_t | \overleftarrow{X}_t) - H(\overrightarrow{X}_t | X_t, \overleftarrow{X}_t). \tag{1}$$

Thus, the PIR may be interpreted as the average information gain, or reduction in uncertainty about the infinite future on learning X_t, given the past. In similar terms, three other information measures can be defined: the entropy rate h_X, the multi-information rate ρ_X [6] and the erasure entropy rate r_X [7], as follows:

$$h_X = H(X_t | \overleftarrow{X}_t), \tag{2}$$

$$\rho_X = I(X_t; \overleftarrow{X}_t) = H(X_t) - H(X_t | \overleftarrow{X}_t), \tag{3}$$

$$r_X = H(X_t | \overleftarrow{X}_t, \overrightarrow{X}_t). \tag{4}$$

Because of the symmetry of the mutual information, the PIR can also be written as $b_X = H(X_t | \overleftarrow{X}_t) - H(X_t | \overrightarrow{X}_t, \overleftarrow{X}_t) = h_X - r_X$. The measures are illustrated in an *information diagram*, or I-diagram [8], in fig. 1, which shows how they partition the marginal entropy $H(X_t)$, the uncertainty about a single observation in isolation; this partitioning is discussed in depth by James *et al* [9].

Dynamic information measures Moving from the general charactarisation of a random process to the analysis of specific sequences, we consider time-varying information measures that can be computed given an unfolding sequence and an assumed process model: from a sequence of observations up to time t, we define two values: (a) the negative log-probability, or *surprisingness* of the observation $X_t{=}x_t$ given the observations so far $\overleftarrow{x}_t \equiv (\ldots, x_{t-1})$,

$$\ell_X^x(t) \triangleq -\log P(X_t{=}x_t | \overleftarrow{X}_t{=}\overleftarrow{x}_t); \tag{5}$$

and (b) the *instantaneous predictive information* (IPI) in the observation $X_t{=}x_t$ about the entire unobserved future $\overrightarrow{X}_t$ given the previous observations $\overleftarrow{X}_t{=}\overleftarrow{x}_t$,

$$i_X^x(t) \triangleq \mathcal{I}(X_t{=}x_t; \overrightarrow{X}_t | \overleftarrow{X}_t{=}\overleftarrow{x}_t), \tag{6}$$

where the conditional information an event about a random variable is defined as the Kullback-Leibler (KL) divergence between the posterior and prior distributions of the variable of interest before and after the event. The terms 'self-information' and 'information content' have also been used for the quantity we have called 'surprisingness'. Before X_t is observed, the the *expected* surprisingness is a measure of the observer's uncertainty about X_t and may be written as an entropy $H(X_t|\overleftarrow{X}_t{=}\overleftarrow{x}_t)$, and the *expected* IPI is the mutual information $I(X_t; \overrightarrow{X}_t|\overleftarrow{X}_t{=}\overleftarrow{x}_t)$ conditioned on the observed past.

3 Predictive information and Bayesian surprise

In this section we examine predictive information in process models with hidden parameters which are initially unknown but gradually inferred from the observations, and demonstrate a connection between Itti and Baldi's 'Bayesian suprise' [10]. Suppose Θ is a random variable representing the unknown parameters of the model and that the observed variables X_t are conditionally iid given Θ, as depicted in fig. 2(a). Thus, the present and future are independent given Θ:

$$I(X_t; \overrightarrow{X}_t | \Theta) = 0. \tag{7}$$

This accounts for the lower zero in the I-diagram of fig. 2(b). Next, we make an additional assumption that, given a long sequence of observations, each additional observation carries less and less extra information about Θ, until, in the limit, any extra observation will not carry any more information about Θ. We call this the zero asymptotic information (ZAI) assumption, and write it as

$$\forall x, \lim_{n\to\infty} \mathcal{I}(X_t{=}x; \Theta | X_{t+1}^{t+n}) = 0, \tag{8}$$

where $X_m^n \equiv (X_m, \ldots, X_n)$. This accounts for the other zero in the I-diagram. Suppose that only a finite segment of the process X_1^{t-1}, has been observed, leaving some uncertainty about Θ, and let O_t denote the observation event $(X_1^t = x_1^t)$. Conditioning on O_{t-1} does not affect the conditional independences given above, and so

$$\mathcal{I}(X_t{=}x_t; \overrightarrow{X}_t | O_{t-1}) = \mathcal{I}(X_t{=}x_x; \Theta | O_{t-1}), \tag{9}$$

that is, the IPI is precisely the Bayesian surprise.

If we relax the assumption that the observations are conditionally independent given the parameters, we find, retaining the ZAI condition, that

$$\mathcal{I}(X_t{=}x_t; \overrightarrow{X}_t | O_{t-1}) = \mathcal{I}(X_t{=}x_t; \overrightarrow{X}_t | \Theta, O_{t-1}) + \mathcal{I}(X_t{=}x_x; \Theta | O_{t-1}). \tag{10}$$

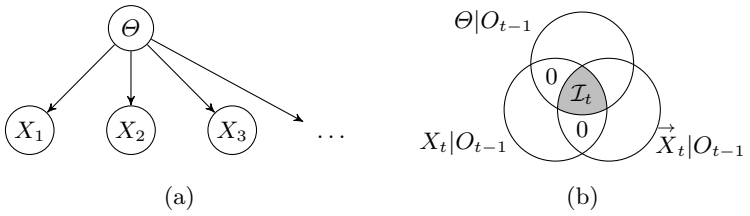

Fig. 2. Surprise and information in an exchangeable random sequence $(X_1, X_2, \ldots)$, which are conditionally independent given the hidden parameters Θ. (a) graphical model representation; (b) I-diagram summarising the situation after observations up to time t. The zeros represent conditional independence assumptions (see main text for details).

Assuming Θ takes values in a set $\mathcal{M}$, the first term on the right-hand side can be expanded as

$$\mathcal{I}(X_t{=}x_t; \overrightarrow{X}_t|\Theta_t, O_{t-1}) = \int_{\mathcal{M}} \mathcal{I}(X_t{=}x_t; \overrightarrow{X}_t|\Theta{=}\theta) p_{\Theta|O_t}(\theta)\, \mathrm{d}\theta, \tag{11}$$

where $p_{\Theta|O_t}$ is the posterior pdf over the parameter space given the observations x_1^t. The second term, the Bayesian surprise is the KL divergence $D(p_{\Theta|O_t}||p_{\Theta|O_{t-1}})$. Thus, we see that the IPI in a system where parameters are being estimated online is composed of two components: the Bayesian surprise, and the IPI for a known parameter value *averaged* over the posterior distribution over parameters.

If, instead of assuming that Θ is constant, we assume it varies slowly, then the above analysis may be taken as an approximation, whose accuracy depends on the extent to which information gained about the parameters is manifested in a finite sequence of future observations corresponding to the time-scale of variation.

4 Process information measures for Gaussian processes

It is known that the entropy rate of a stationary Gaussian process can be expressed in terms of its power spectral density (PSD) function $S : \mathbb{R} \to \mathbb{R}$, which is defined as the discrete-time Fourier transform of the autocovariance sequence $\gamma_k = \mathrm{E}\, X_t X_{t-k}$, where E is the expectation operator. For a Gaussian process, the entropy rate is the Kolmogorov-Sinai entropy:

$$h_X = \frac{1}{2}\left(\log(2\pi e) + \frac{1}{2\pi}\int_{-\pi}^{\pi} \log S(\omega)\, \mathrm{d}\omega\right). \tag{12}$$

Dubnov [6] gave the multi-information rate (MIR) of a stationary Gaussian process in terms of the spectral density $S(\omega)$ as:

$$\rho_X = \frac{1}{2}\left(\log\left[\frac{1}{2\pi}\int_{-\pi}^{\pi} S(\omega)\, \mathrm{d}\omega\right] - \frac{1}{2\pi}\int_{-\pi}^{\pi} \log S(\omega)\, \mathrm{d}\omega\right), \tag{13}$$

which follows from the observation that $H(X_t) = \log(2\pi e \gamma_0)$ and the relation $\rho_X = H(X_t) - h_X$. Verdú and Weissman [7] give a general expression for the

erasure entropy rate of a Gaussian process in terms of its power spectral density. Using this and writing the entropy rate in a slightly different form, we obtain

$$b_X = \frac{1}{2}\left(\log\left[\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{1}{S(\omega)}\,\mathrm{d}\omega\right] - \frac{1}{2\pi}\int_{-\pi}^{\pi}\log\frac{1}{S(\omega)}\,\mathrm{d}\omega\right), \tag{14}$$

which, compared with the expression (13), suggests a duality between the multi-information and predictive information rates on the one hand, and Gaussian processes whose power spectra are mutually inverse on the other. A similar duality was noted by [5] in relation to the multi-information and the binding information (the extensive counterpart to the predictive information rate) in finite sets of discrete-valued random variables.

Autoregressive Gaussian processes An autoregressive Gaussian process of order N is a real-valued random process such that $X_t = U_t - \sum_{k=1}^{N} a_k X_{t-k}$, where the *innovations* U_t are iid Gaussian random variables with zero mean and variance σ^2, and the a_k are the autogressive or prediction coefficients. The class of such processes is known as AR(N). If the coefficients a_k are such that the filter is stable, the process will be stationary and thus may have well defined entropy and predictive information rates. It is relatively straighforward to show that the entropy and predictive information rates of an AR(N) process are

$$h_X = \tfrac{1}{2}\log(2\pi e\sigma^2), \qquad b_X = \tfrac{1}{2}\log\left(1 + \sum_{k=1}^{N} a_k^2\right). \tag{15}$$

The multi-information rate ρ_X does not have a simple general expression in terms of the parameters and can be computed either by solving the Yule-Walker equations to get the marginal entropy or from the power spectrum.

Adding noise to avoid infinite information If no restrictions are placed on the PSD, both ρ_X and b_X are unbounded. The reason for this, we suggest, lies in the assumption that the real-valued random variables can be observed with *infinite* precision. This rather un-physical situation can be remedied if we introduce noise, observing the process X through a noisy channel Y, where $Y_t = X_t + V_t$ and $(V_t)_{t\in\mathbb{Z}}$ is white noise. In this case, each observation Y_t can only yield a finite amount of information about X_t. For AR(N) processes, this results in an inverted-'U' relationship between the PIR and both the multi-information and entropy rates, with finite maxima for all information measures.

Dynamic information measures Since X_t is conditionally Gaussian, the dynamic surprisingness measure $\ell_X^x(t)$ defined earlier (5) is a function of the deviation of x_t from its expected value $\hat{x}_t = \mathrm{E}\left(X_t | \overleftarrow{X}_t = \overleftarrow{x}_t\right)$, which, for an autoregressive process, can be computed directly from the prediction coefficients and the previous observations. The result can be written as (see [11] for details)

$$\ell_X^x(t) = h_X + \pi e^{1-2h_X}(x_t - \hat{x}_t)^2 - \tfrac{1}{2}. \tag{16}$$

Note that $(x_t - \hat{x}_t)$ is by construction the *innovation* at time t and thus, the expectation of $(x_t - \hat{x}_t)^2$ is $e^{2h_X}/2\pi e$ *independently* of t, which means that at all times, expectation of $\ell_X^x(t)$ is constant at h_X. It also means that the sequence of surprisingness values is itself uncorrelated in time. This is in marked constrast with the situation for Markov chains [4], where, in general, the expected suprise depends on the previous observation and thus varies in time, reflecting the observer's varying levels of uncertainty about the next observation. In a Gaussian processes, this predictive uncertainty is constant and therefore does not provide any useful structural analysis of the sequence. The IPI (6) can be expressed in several ways, but perhaps the most illuminating (see [11] for a derivation) is

$$i_X^x(t) = \left[1 - e^{-2b_X}\right] \left[\ell_X^x(t) - h_X\right] + b_X. \tag{17}$$

Since h_X is the expectation of $\ell_X^x(t)$ and b_X is the expectation of $i_X^x(t)$, this has a rather perspicacious reading: *the deviations of the surprisingness and the IPI from their expectations are proportional to one another*. The constant of proportionality varies from zero when $b_X = 0$ to 1 as $b_X \to \infty$. As with the expected surprisingness, the expected IPI is constant and equal to b_X.

4.1 AR estimation and Bayesian suprise

Our method for spectral estimation is based on Kitagawa and Gersch's [12] 'spectral smoothness prior'—they consider autoregressive Gaussian processes and introduce a measure of spectral smoothness to be used as a regulariser in spectral estimation when the model order is high but the amount of data available is low. They show how this leads to a Gaussian prior with independent coefficients such that $a_k \sim \mathcal{N}(0, \lambda^{-2}k^{-2\alpha})$, where $\alpha > 0$ controls the order of smoothness favoured and λ controls the overall strength of the prior. This is especially convenient since, when parameterised by the a_k, the (multivariate) Gaussian is a conjugate prior, so that the posterior distribution remains Gaussian as data accumulates.

We adapted Kitagawa and Gersch's offline method to online estimation of both the innovation variance σ^2 and the coefficients $a_{1:N}$ using a conjugate prior, which is inverse-Gamma for σ^2 and conditionally Gaussian for $a_{1:N}$. At time t, the posterior is represented by its natural parameters η_t (in language of exponential families), which are essentially the sufficient statistics of the data with respect to the model. This amounts to keeping a running estimate of the autocovariance of the signal at lags from zero to N. In order to allow for slow variations in the spectrum, a forgetting factor is included, resulting in an exponentially decaying memory of older observations. The recursive update can be written as $\eta'_{t-1} = (\tau - 1/\tau)\eta_{t-1}, \eta_t = T(x_t; x_{t-N}^{t-1}) + \eta'_{t-1}$, where τ is the effective time constant and $T(\cdot)$ computes the sufficient statistics for the current observation given the previous N. The initial state η_0 is derived from the spectral smoothness prior. Given η_t, the Bayesian surprise is the KL divergence between the two distributions specified by η_t and η'_{t-1}, which we can write as $D_\eta(\eta_t||\eta'_{t-1})$. The entropy rate and PIR of the currently estimated process are computed from the posterior mean of $a_{1:N}$ computed from η_t. Finally, the marginal variance

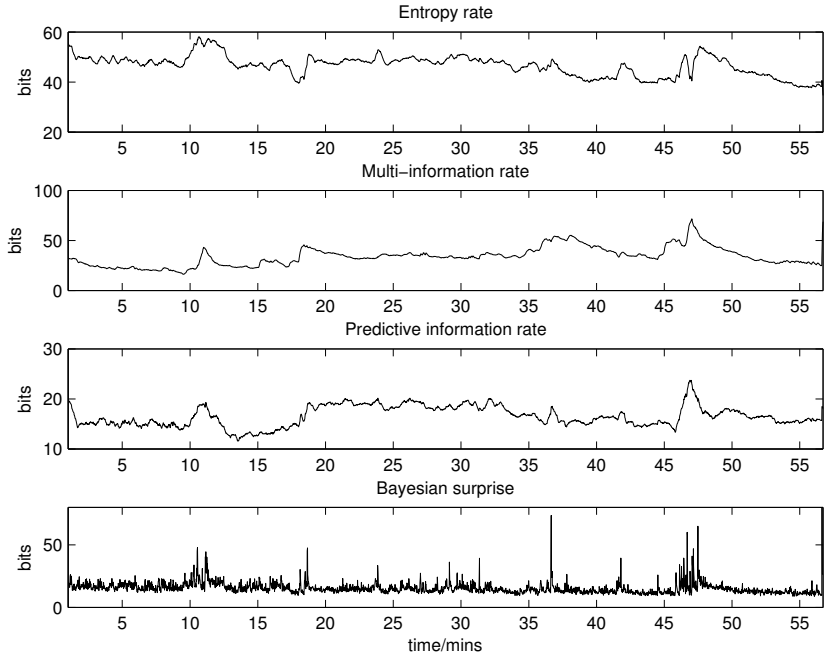

Fig. 3. An analysis of Steve Reich's *Drumming* in terms of process information measures. The spikes in the Bayesian surprise correspond to significant events in the score (changes in instrumentation), while the traces of features of the predictive information and entropy rates can be related to structural features of the music. Part boundaries are at around 18, 36, and 46 minutes.

and thus the marginal entropy $H(X_t)$ are estimated directly from the signal in order compute the MIR as $\rho_X^{(t)} = H(X_t) - h_X^{(t)}$. This was found to be more stable numerically than computing the MIR from $a_{1:N}$, since the estimated coefficients would sometimes yield an unstable filter with an undefined MIR.

5 Applications to music analysis

We applied the above methods to a recording of Steve Reich's *Drumming*, following the general approach of [13]: the signal was represented as a sequence of short-term Mel-frequency spectra (256 bands, frame length 186 ms, hop size 46 ms, frame rate approx. 21 Hz); the first 32 decorrelated principal components were computed offline. Then, treating each channel independently, dynamic mean subtraction (time constant about 10 mins) was followed by online spectral estimation using an AR(24) model, with a forgetting time constant of about 12 s, $\alpha = 1$ and $\lambda = 1$. The resulting information measures were summed across all 32 channels to produce the results illustrated in fig. 3. Part boundaries and changes in instrumentation are well captured by peaks in the Bayesian surprise.

6 Discussion and conclusions

The PIR and IPI were found to be simply expressible for stationary discrete-time Gaussian processes, with a certain duality between the PIR and MIR with respect to spectral inversion (exchaninging poles for zeros). The expressions for dynamic surprise and instantenous predictive information suggest that stationary Gaussian processes are relatively lacking in temporal structure. The identification of the Bayesian surprise as a component of the IPI when learning parameterised models links the two activities of learning about parameters and gaining new information about future observations. The accuracy of these results when used as an approximation for models with time-varying parameters will depend on the information geometry of the model and will be a subject of future work.

When appplied to the analysis of a recording of Steve Reich's *Drumming*, the information measures were found to vary systematically across the piece, with several structural boundaries and features visible. As we chose to use a framework modelled on that of [13], a fuller analysis and comparison with Dubnov's multi-information rate analysis will be the subject of future work.

References

1. Crutchfield, J., Packard, N.: Symbolic dynamics of noisy chaos. Physica D: Nonlinear Phenomena **7** (1983) 201–223
2. Grassberger, P.: Toward a quantitative theory of self-generated complexity. International Journal of Theoretical Physics **25** (1986) 907–938
3. Bialek, W., Nemenman, I., Tishby, N.: Predictability, complexity, and learning. Neural Computation **13** (2001) 2409–2463
4. Abdallah, S.A., Plumbley, M.D.: Information dynamics: Patterns of expectation and surprise in the perception of music. Connection Science **21** (2009) 89–117
5. Abdallah, S.A., Plumbley, M.D.: A measure of statistical complexity based on predictive information with application to finite spin systems. Physics Letters A **376** (2012) 275 – 281
6. Dubnov, S.: Spectral anticipations. Computer Music Journal **30** (2006) 63–83
7. Verdú, S., Weissman, T.: Erasure entropy. In: IEEE International Symposium on Information Theory (ISIT 2006). (2006) 98–102
8. Yeung, R.: A new outlook on Shannon's information measures. Information Theory, IEEE Transactions on **37** (1991) 466–474
9. James, R.G., Ellison, C.J., Crutchfield, J.P.: Anatomy of a bit: Information in a time series observation. Chaos **21** (2011) 037109
10. Itti, L., Baldi, P.: Bayesian surprise attracts human attention. In: Advances Neural in Information Processing Systems (NIPS 2005). Volume 19., Cambridge, MA, MIT Press (2005) 547–554
11. Abdallah, S.A., Plumbley, M.D.: Instantaneous predictive information in Gaussian processes. Unpublished technical note. (2012)
12. Kitagawa, G., Gersch, W.: A smoothness priors time-varying ar coefficient modeling of nonstationary covariance time series. IEEE Transactions on Automatic Control **30** (1985) 48–56
13. Dubnov, S., McAdams, S., Reynolds, R.: Structural and affective aspects of music from statistical audio signal analysis. Journal of the American Society for Information Science and Technology **57** (2006) 1526–1536