

RESEARCH ARTICLE

Information Dynamics: Patterns of expectation and surprise in the perception of music

Samer Abdallah and Mark Plumbley

Centre for Digital Music, Queen Mary, University of London,
Mile End Road, London E1 4NS

(Received 00 Month 200x; final version received 00 Month 200x)

Measures such as entropy and mutual information can be used to characterise random processes. In this paper, we propose the use of several *time-varying* information measures, computed in the context of a probabilistic model which evolves as a sample of the process unfolds, as a way to characterise temporal structure in music. One such measure is a novel *predictive information rate* which we conjecture may provide an explanation for the ‘inverted-U’ relationship often found between simple measures of randomness (e.g. entropy rate) and judgements of aesthetic value (1). We explore these ideas in the context of Markov chains using both artificially generated sequences and two pieces of minimalist music by Philip Glass, showing that even such a manifestly simplistic model (the Markov chain), when interpreted according to information dynamic principles, produces a structural analysis which largely agrees with that of an expert human listener.

Keywords: information theory; expectation; surprise; subjective probability; Markov chain; music.

1. Expectation and surprise in music

One of the more salient effects of listening to music is to create *expectations* of what is to come next, which may be fulfilled immediately, after some delay, or not at all as the case may be. This is the thesis put forward by, amongst others, music theorists L. B. Meyer (2) and Narmour (3). In fact, this insight predates Meyer quite considerably; for example, it was elegantly put by Hanslick (4) in the nineteenth century:

‘The most important factor in the mental process which accompanies the act of listening to music, and which converts it to a source of pleasure, is frequently overlooked. We here refer to the intellectual satisfaction which the listener derives from continually following and anticipating the composer’s intentions—now, to see his expectations fulfilled, and now, to find himself agreeably mistaken. It is a matter of course that this intellectual flux and reflux, this perpetual giving and receiving takes place unconsciously, and with the rapidity of lightning-flashes.’

An essential aspect of this is that music is experienced as a phenomenon that ‘unfolds’ in time, rather than being apprehended as a static object presented in its entirety. Meyer argued that musical experience depends on how we change and revise our conceptions *as events happen*, on how expectation and prediction interact with occurrence, and that, to a large degree, the way to understand the effect of music is to focus on this ‘kinetics’ of expectation and surprise.

Corresponding author; email: samer.abdallah@elec.qmul.ac.uk

The business of making predictions and assessing surprise is essentially one of reasoning under conditions of uncertainty and manipulating degrees of belief about the various proposition which may or may not hold, and, as has been argued elsewhere (5, 6), best quantified in terms of Bayesian probability theory. Thus, we suppose that when we listen to music, expectations are created on the basis of our familiarity with various stylistic norms using models that encode the statistics of music in general, the particular styles of music that seem best to fit the piece we happen to be listening to, and the emerging structures peculiar to the current piece. There is experimental evidence that human listeners are able to internalise statistical knowledge about musical structure, e.g. (7, 8), and also that statistical models can form an effective basis for computational analysis of music, e.g. (9–11).

1.1. *Music and information theory*

Given a probabilistic framework for music modelling and prediction, it is a small step to apply quantitative information theory (12) to the models at hand. The relationship between information theory and music and art in general has been the subject of some interest since the 50s (e.g. (2, 13, 14)). The general thesis is that perceptible qualities and subjective states like uncertainty, surprise, complexity, tension, and interestingness are closely related to information-theoretic quantities like entropy, relative entropy, and mutual information. Berlyne (1) called such quantities ‘collative variables’, since they are to do with patterns of occurrence rather than medium-specific details. and developed the ideas of ‘information aesthetics’ in an experimental setting.

Previous work in this area treated the various information theoretic quantities such as entropy as if they were intrinsic properties of the stimulus—subjects were presented with a sequence of tones with ‘high entropy’, or a visual pattern with ‘low entropy’. These values were determined from some known ‘objective’ probability model of the stimuli, or from simple statistical analyses such as computing empirical distributions. Our approach is explicitly to consider the role of the observer in perception, and more specifically, to consider estimates of entropy etc. with respect to *subjective* probabilities.

Bringing these strands together, our working hypothesis is that as we humans listen to a piece of music, we maintain a dynamically evolving statistical model that enables us to make predictions about how the piece will continue, relying on both our previous experience of music and the immediate context of the piece. As events unfold, we revise our model and hence our probabilistic belief state, which includes predictive distributions over future observations. These distributions and changes in distributions can be characterised in terms of a handful of information theoretic-measures such as entropy and relative entropy. By tracing the evolution of a these measures, we obtain a representation which captures much of the significant structure of the music.

This approach has a number of features. *Abstraction*: Because it is sensitive mainly to *patterns* of occurrence, rather the details of which specific things occur, it operates at a level of abstraction removed from the details of the sensory experience and the medium through which it was received, suggesting that the same approach could, in principle, be used to analyse and compare information flow in different temporal media regardless of whether they are auditory, visual or otherwise. *Generality*: This approach does not proscribe which probabilistic models should be used—the choice can be guided by standard model selection criteria such as Bayes factors (15), etc. *Richness*: In particular, it may be effective to use a model with time-dependent latent variables, such as a hidden Markov model. In

these cases, we can track changes in beliefs about the hidden variables as well as the observed ones, adding another layer of richness to the description while maintaining the same level of abstraction. For example, harmony (i.e., the ‘current chord’) in music is not stated explicitly, but rather must be inferred from the musical surface; nonetheless, a sense of harmonic progression is an important aspect of many styles of music. *Subjectivity*: Since the analysis is dependent on the probability model the observer brings to the problem, which may depend on prior experience or other factors, and which may change over time, inter-subject variability and variation in subjects’ responses over time are fundamental to the theory. It is essentially a theory of subjective response and does not seek or require the possibility of an objective analysis of music.

The remainder of the paper is organised as follows: in §2 we provide general definitions of the information measures that we are going to examine; in §3 we show how these measures can be computed for a particular model, the Markov chain, and examine the information dynamics of sequences generated artificially from known Markov chains; in §4 we examine the information dynamics of online learning in Markov chains. We apply the Markov chain model to minimalist music by Philip Glass §5, and show how the information-dynamic approach yields a plausible structural analysis that largely agrees of a human expert listener. In §6 we are in a position to relate our approach with previous work in the same area. We wrap-up with discussion and conclusions in §7.

2. Model-based observation of random processes

In this section we define some of the information measures that a model-based observer can compute given a realisation of a random process and a statistical model that can be updated dynamically as the process unfolds. This observer-centric view highlights the point that the probabilities we consider here are essentially *subjective* probabilities, and do not require any ‘objective’ or frequentist interpretation. The observer’s model need not be the ‘correct’ one, and we need not rely on the epistemologically questionable notion of a ‘correct’ model existing (6).

Consider a snapshot of a stationary random process taken at a certain time: we can divide the timeline into an infinite ‘past’ and ‘future’, and a notional ‘present’ of finite duration. Observations of the process can be grouped into three random variables, say Z , Y , and X , corresponding to these three time intervals respectively (see fig. 1). The model is summarised by the observer’s probability distribution $p_{XY|Z}$ over the present and future given the past. For discrete variables, $p_{XY|Z}(x, y|z)$ is the probability with which the observer expects to see x followed by y given that it has already seen z . We can now consider how the observer’s belief state evolves when it learns that $X=x$.

2.1. ‘Surprise’-based measures

To obtain a first set of four information measures, we marginalise out the future Y to get the distribution for the immediate prediction, $p_{X|Z}$. The negative log-probability

$$\mathcal{L}(x|z) \triangleq -\log p_{X|Z}(x|z), \quad (1)$$

can be thought of as the ‘surprisingness’ of x in the context of z . The expectation of this quantity (given a particular z) is the entropy of the predictive distribution,

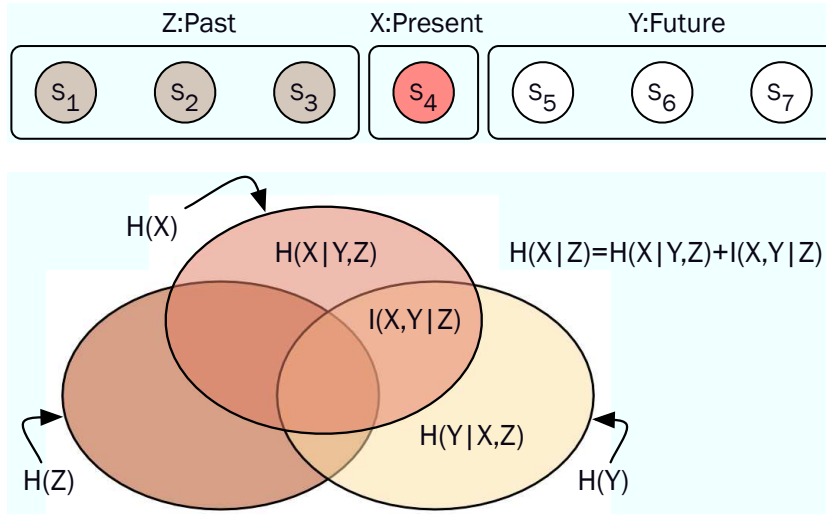


Figure 1. By grouping the elements of a random sequence into a *past*, *present*, and *future*, we can consider a number of information measures, some of which are well known, like the entropy rate $H(X|Z)$, and some of which have not, to our knowledge, been investigated before, such as the average predictive information rate $I(X, Y|Z)$. The relationship between several such interest can be visualised as a Venn diagram. Note that Z and Y actually stand for the *infinite* past and future and are only shown as finite for visualisation purposes.

which we will write as $H(X|Z=z)$ to emphasise that it is a function of the observed past z ; it is a measure of the observer’s uncertainty about X before the observation is made.

Once the observer sees that $X=x$, it can compute its surprisingness $\mathcal{L}(x|z)$, but for some classes of model it may be possible to average $\mathcal{L}(x|z)$ over the past contexts (given the current model) that could have lead to the current observation, that is, over $Z|X=x$. This average in-context surprisingness of the symbol x might be useful as a sort of static analysis of the model, helping to pick out which are the most significant states in the state space. By averaging $\mathcal{L}(x|z)$ over *both* variables, we obtain the entropy rate $H(X|Z)$ of the process according to the observer’s current model. Thus, the first four measures are the surprisingness and its three averages over $(X|Z=z)$, $(Z|X=x)$, and (X, Z) jointly.

2.2. Predictive information-based measures

Perhaps more important than intrinsic surprisingness of an observation is the information it carries *about* the unobserved future, *given* that we already know the past. This is what we are calling the *predictive information* (PI). Hence, to obtain a second set of four information measures, we consider the information supplied about Y by the observation that $X=x$, given that we already know $Z=z$, quantified as the Kullback-Leibler (KL) divergence between the predictive distribution over Y before and after the event $X=x$, that is,

$$\mathcal{I}(x|z) \triangleq I(X=x, Y|Z=z) = D(p_{Y|X=x, Z=z} || p_{Y|Z=z}), \quad (2)$$

where $p_{Y|Z=z}(y) = \int p_{XY|Z=z}(x, y) dx$ and $D(\cdot||\cdot)$ is the KL divergence between two distributions. Like $\mathcal{L}(x|z)$, this is a function of the observations z and x , and we can take expectations over X or Z or both. Averaging over the prediction $X|Z=z$, that is, computing $E_{X|Z=z} \mathcal{I}(X|z)$, tells us the amount of new information we *expect* to receive from the next observation about the future. It could be useful as a guide to how much attention needs to be directed towards the next event even before

it happens. This is different from Itti and Baldi’s proposal that Bayesian *surprise* attracts attention (16), as it is a mechanism which can operate *before* the surprise occurs.

The average of the PI over preceding contexts $Z|X=x$, that is, the expectation $E_{Z|X=x} \mathcal{I}(x|Z)$, is the amount of information about the future carried, on average, by each value in the state space of X . As before, this tells us something about the significance of each symbol in the alphabet, picking out which symbols tend to be most informative about the future. One might predict that these states will tend to appear as ‘onset’ states, or as the ‘foreground’ against a ‘background’ of the states that tend not to carry much information.

Averaging over both X and Z gives us the *predictive information rate* (PIR), which is, for a given random process model, the average rate at which new information arrives about the future. The expression reduces to what one might call a ‘conditional mutual information’ (see fig. 1):

$$I(X, Y|Z) = H(Y|Z) - H(Y|X, Z). \quad (3)$$

Overall, the four measures in the second set are $\mathcal{I}(x|z)$ and its expectations over X , Z , and (X, Z) jointly. Unlike those in the first set, these measures are computed in terms of KL divergences and hence are invariant to invertible transformations of the observation spaces: the random process could be ‘transcoded’ using different symbols and perhaps a different modality, and as long as the transcoding was invertible, the predictive information measures would remain the same.

2.3. Information about model parameters

Finally, another information measure can be obtained by considering an observer using an explicitly parameterised model. In this case, the observer’s belief state would include a probability distribution for the parameters Θ . Each observation would cause a revision of that belief state and hence supply information about the parameters, which we will again quantify as the KL divergence between prior and posterior distributions $D(p_{\Theta|X=x, Z=z} || p_{\Theta|Z=z})$. We call this the ‘model information rate’.

Note that in a rigorous analysis of the predictive information in a model which includes unknown parameters, information gained about the parameters would also show up as information gained about future observations, since the correct way to compute the probability of future observations in these models is to take account of uncertainty about the parameters and integrate them out, though in most cases this computation will be intractable.

2.4. Predictive information rate as a measure of structure

Many studies looking into the relationship between stochastic complexity as measured by entropy or entropy rate, and what has variously been called ‘pleasingness’, ‘hedonic values’, and ‘aesthetic value’, reveal an inverted ‘U’ shaped curve (see fig. 2) where the highest value is attached to processes of intermediate entropy (1).

This type of relationship (though not in quantitative information-theoretic terms) was also observed by Wundt (17). Intuitively, patterns which are too deterministic and ordered are boring, while those which are too random are perceived as unstructured, featureless, and, in a sense, ‘uniform’, in the way that white noise is. Hence, a sequence can be uninteresting in two opposite ways: by being utterly predictable *or* by being utterly unpredictable. Meyer (18) hints at the same thing

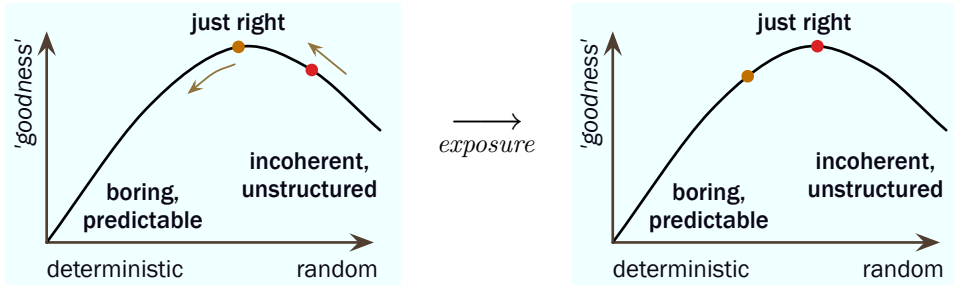


Figure 2. Relationship between apparent complexity and aesthetic value, and the change in value judgement which sometimes occurs after prolonged exposure.

while discussing the relation between the rate of information flow and aesthetic experience, suggesting that ‘If the amount of information [by which he means entropy and surprisingness] is inordinately increased, the result is a kind of cognitive white noise.’

The explanations for this usually appeal to a need for a ‘balance’ between order and chaos, unity and diversity, and so on, in a generally imprecise way. However, the predictive information rate (3) seems to incorporate this balance automatically (see fig. 3), achieving a maximum for sequences which are neither deterministic nor totally uncorrelated across time. Our interpretation of this is that when each event appears to carry no new information about the unknown future, it is not worth attending to, and in a way, meaningless. As we have a quantitative prescription for computing the PIR, this is something that could be tested experimentally with human subjects.

Berlyne (1, ch. 13) also discusses the effect which repeated exposure to a stimulus has on its perceived aesthetic value. The evidence he reviews is conflicting: in some cases repeated exposure leads to an increase in preference, while in others, to a decrease. Berlyne argues that this can be understood as a process of migration to the left along the Wundt curve as the subjective complexity of the stimuli decreases due to the observer learning something of its structure, as shown in fig. 2. Stimuli starting out on the right of the curve will be liked more as the observer becomes more familiar with them, while those starting near the top or on the left will be liked less and less. We will return to this in §4 where we show that the predictive information rate can display similar behaviour when computed using a probability model that adapts over time.

3. Information dynamics in Markov chains

To illustrate the how the measures defined in §2 can be computed in practice, we will consider one of the simplest random processes, a first order Markov chain. Let S be a Markov chain with a finite state space $\{1, \dots, N\}$ such that S_t is the random variable representing the t^{th} element of the sequence. The model is parameterised by a transition matrix $a \in \mathbb{R}^{N \times N}$ encoding the distribution of any element of the sequence given previous one, that is $p(S_{t+1}=i|S_t=j) = a_{ij}$. Since we require the process to be stationary, we set the distribution for the initial element S_1 to the equilibrium distribution of the transition matrix, that is, $p(S_1=i) = \pi_i^a$ where π^a is a column vector satisfying $a\pi^a = \pi^a$. To ensure that the equilibrium distribution is unique, we also require that the Markov chain be irreducible. Under these conditions, the Markov chain will have an entropy rate which can be written

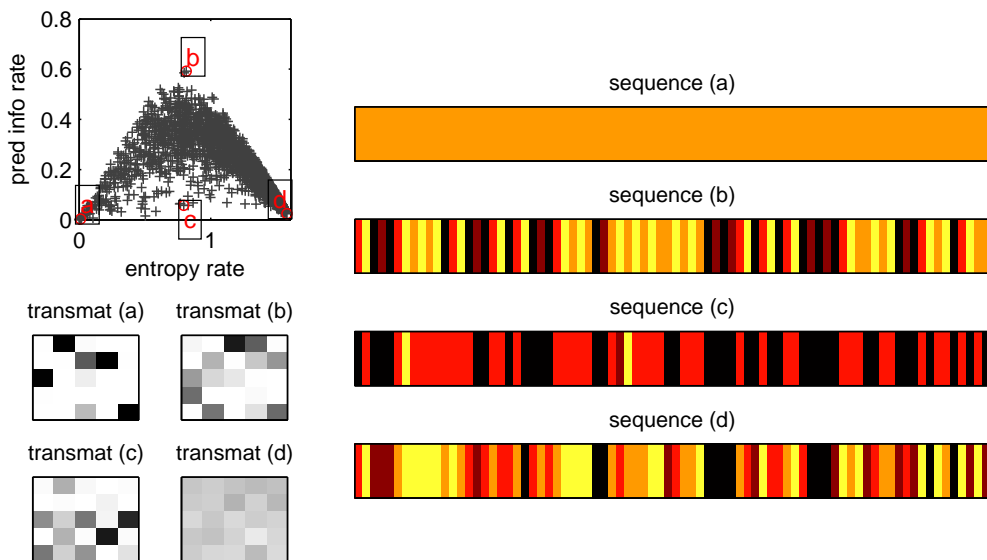


Figure 3. The space of transition matrices explored by generating them at random and plotting entropy rate vs PIR. (Note inverted ‘U’ relationship). Four of the transition matrices are shown along with sample sequences. Sequence (a) is simply the endless repetition of state 4. Matrix (d) is almost uniform. Matrix (b) has the highest PIR.

as a function of a alone:

$$\dot{\mathcal{H}} : \mathbb{R}^{N \times N} \rightarrow \mathbb{R}, \quad \dot{\mathcal{H}}(a) = \sum_{i=1}^N \pi_i^a \sum_{j=1}^N -a_{ji} \log a_{ji}. \quad (4)$$

The Markov dependency structure means that, for the purposes of computing the measures defined in § 2, the ‘past’ and ‘future’ at time t can be collapsed down to the previous and next elements of the chain (see appendix). In terms of our earlier notation, we can set $Z = S_{t-1}$, $X = S_t$, and $Y = S_{t+1}$. Equations (6) and (7) below give expressions for the eight information measures from the first two sets defined in § 2. Some of these are expressed in terms of the ‘time-reversed’ transition matrix defined as

$$a_{ij}^\dagger = p(S_{t-1}=j | S_t=i) = a_{ij} \pi_j^a / \pi_i^a. \quad (5)$$

Note that the over- and under-bars are intended as mnemonics for the expectations over S_t and S_{t-1} respectively.

The first four, ‘surprise’-based, measures are

$$\begin{aligned} \mathcal{L}(i|j) &= -\log p(S_t=j | S_{t-1}=i) = -\log a_{ij}, \\ \bar{\mathcal{L}}(j) &= \mathbb{E}_{i \sim S_t | S_{t-1}=j} \mathcal{L}(i|j) = \sum_{i=1}^N a_{ij} \mathcal{L}(i|j), \\ \underline{\mathcal{L}}(i) &= \mathbb{E}_{j \sim S_{t-1} | S_t=i} \mathcal{L}(i|j) = \sum_{j=1}^N a_{ij}^\dagger \mathcal{L}(i|j), \\ \underline{\bar{\mathcal{L}}} &= H(S_{t+1} | S_t) = \dot{\mathcal{H}}(a). \end{aligned} \quad (6)$$

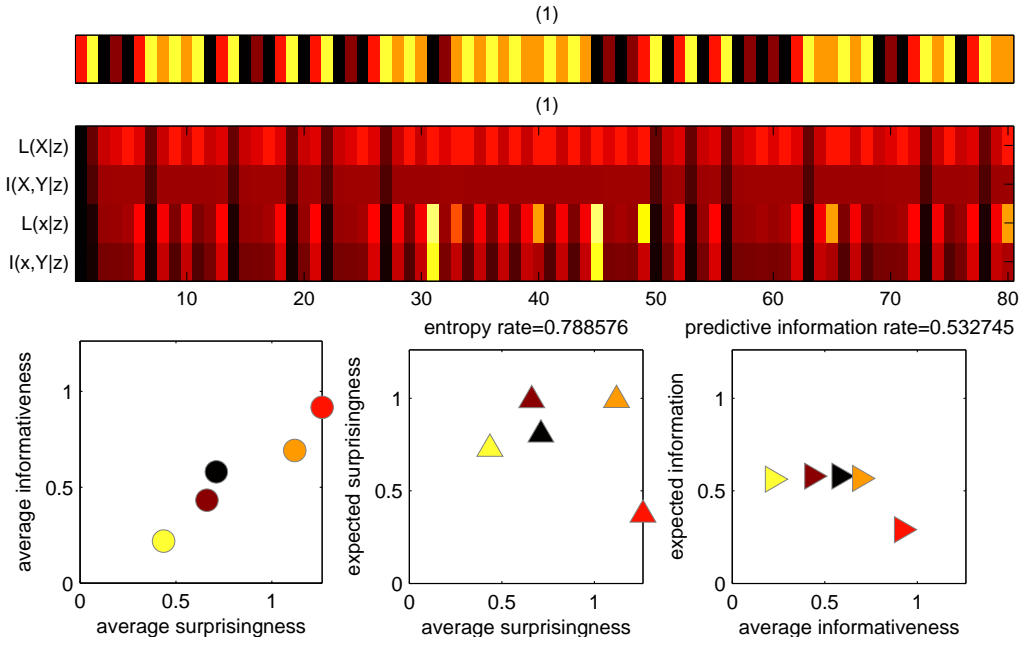


Figure 4. Analysis of transition matrix and sequence (b) from fig. 3. The upper panels show the sequence itself along with the dynamic evolution of, from top to bottom, the predictive uncertainty $\overline{\mathcal{L}}(j)$, the expected predictive information $\overline{\mathcal{I}}(j)$, the surprisingness $\mathcal{L}(i|j)$ and the predictive information $\mathcal{I}(i|j)$, where i and j stand for the current and previous symbols respectively. The lower panels summarise the static analysis of the system, plotting the average surprisingness $\underline{\mathcal{L}}(i)$, the average informativeness $\underline{\mathcal{I}}(i)$, the predictive uncertainty $\overline{\mathcal{L}}(i)$, and the information expectancy $\overline{\mathcal{I}}(i)$ generated by each of the five states.

The second four, predictive-information-based, measures are

$$\begin{aligned}
 \mathcal{I}(i|j) &= D(p_{S_{t+1}|S_t=i} || p_{S_{t+1}|S_t=j}) = \sum_{k=1}^N a_{ki} (\log a_{ki} - \log [a^2]_{kj}), \\
 \overline{\mathcal{I}}(j) &= \mathbb{E}_{i \sim S_t | S_{t-1}=j} \mathcal{I}(i|j) = \sum_{i=1}^N a_{ij} \mathcal{I}(i|j), \\
 \underline{\mathcal{I}}(i) &= \mathbb{E}_{j \sim S_{t-1} | S_t=i} \mathcal{I}(i|j) = \sum_{j=1}^N a_{ij}^\dagger \mathcal{I}(i|j), \\
 \underline{\overline{\mathcal{I}}} &= I(S_t, S_{t+1} | S_{t-1}) = \dot{\mathcal{H}}(a^2) - \dot{\mathcal{H}}(a).
 \end{aligned} \tag{7}$$

3.0.1. Relationship between entropy rate and predictive information rate

For a given size of state space N , the entropy rate can vary between zero for a deterministic sequence and $\log N$ for an uncorrelated sequence with $a_{ij} = 1/N$ for all i, j . Between these extremes, we find that the Markov chains that maximise the PIR have intermediate entropy. The scatter plot in fig. 3 was obtained by generating transition matrices at random by drawing each column independently from a Dirichlet distribution. We also investigated optimising the PIR directly using a general purpose optimiser. We found that, for a range of different N , relatively sparse transition matrices maximise the PIR (see fig. 5). The 16×16 transition matrix is typical of what happens as N is increased: the conditional distribution for each antecedent state is approximately uniformly distributed across 3 or 4 states.

4. Subjective information and model mismatch

In preceding analysis, the surprisingness and predictive information were computed with respect to the observer's probabilistic model on the understanding that

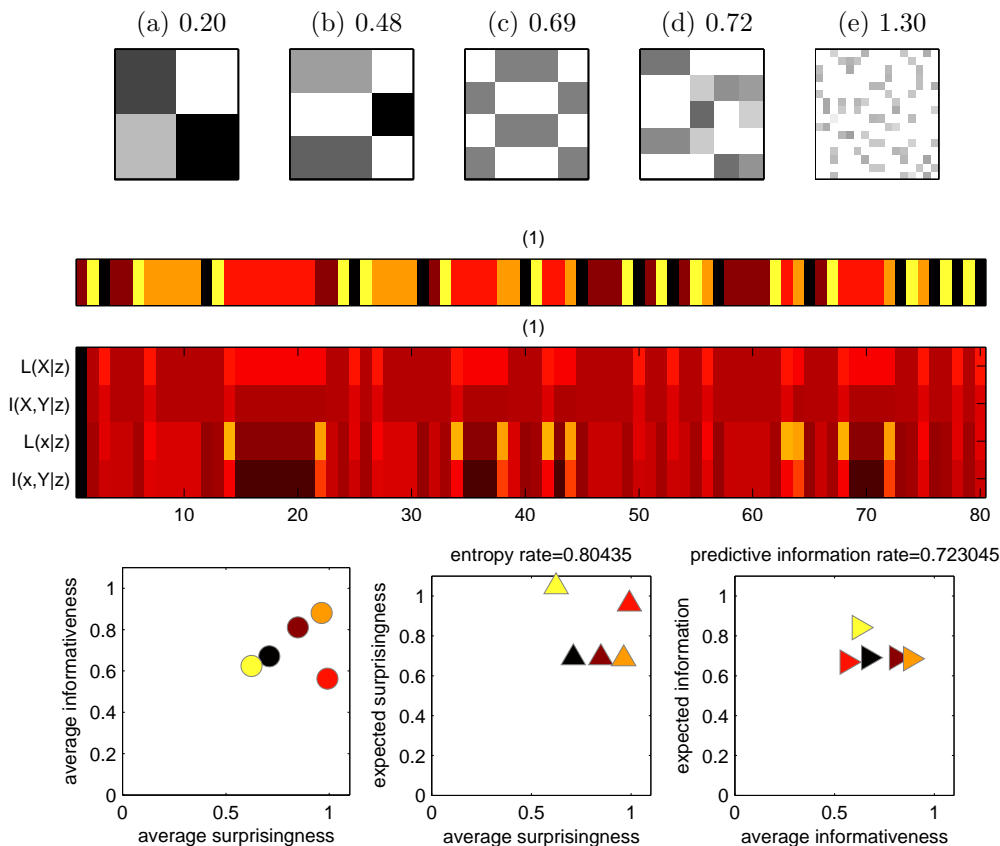


Figure 5. The results of direct numerical optimisation of the PIR for different state space sizes N . The number over each transition matrix is its PIR in nats/symbol ($1 \text{ nat} = \log_2 e \approx 1.44 \text{ bits}$). The panels below show a sample from transition matrix (d) and an information dynamic analysis as in fig. 4

they represent the observer’s subjective surprise and changes in beliefs on observing each symbol. The various averages used to obtain the entropy rate and predictive information rate were taken using the distributions implied by the model itself. When applied to some given sequence observations, these theoretical averages may or may not be close to the empirical average levels of surprise and information gain experienced by the observer as it processes the sequence. This will depend on whether or not the observer’s model is a ‘good’ model of the data.

In some cases, such as when data is generated explicitly by sampling from some particular distribution, there can be said to be a ‘true’ model to which the observer’s model can be compared. In others, the existence of a ‘true’ model may be questionable, or at least, not verifiable in practice. The most we can say in such cases is that one model may or may not be better than another according to certain criteria such as those advocated by Bayesians (15). In these cases, we take de Finetti at his word and say ‘there are no real [i.e. objective] probabilities’, only subjective ones (19).

However, returning to the Markov chain model we analysed in §3, we can ask, what are the average levels of surprise etc. experienced when an observer using one Markov transition matrix processes a sequence generated from a Markov chain using another transition matrix? That is, what if some or all of the averaging operations in equations (6) and (7) are carried out with respect to the true generative distribution rather than the observer’s model? We will see that this leads to several variants of what was the entropy rate and predictive information rate, depending on how the averages are taken. In the following, we will assume an ergodic system

so that we can equate these ensemble averages with the time averages that an observer could estimate given a long enough sample.

4.1. *Surprise-based quantities*

We continue to denote the observer's probabilities with a p , but now introduce the generative probabilities with a q , e.g., $q_{X|Z}$, $q_{Y|X,Z}$ etc.. To parameterise the generative Markov chain, we introduce the generative transition matrix g .

The surprisingness of the observation ($X = x|Z = z$) and the information it provides about the future are the same as in equations 1 and 2, since these are defined entirely in terms of the observer's subjective probability distributions. However, the average surprise following the observation $Z = z$ can now be computed using the actual distribution of symbols that occur after z as well as the distribution that the observer expects. Similarly, the average in-context surprisingness of symbol x as it actually occurs can be computed using the true distribution $q_{Z|X=x}$. The general expressions for these averages are

$$\begin{aligned}\bar{\mathcal{L}}^*(z) &= \sum_x \mathcal{L}(x|z)q(x|z), \\ \underline{\mathcal{L}}_*(z) &= \sum_z \mathcal{L}(x|z)q(z|x),\end{aligned}\tag{8}$$

Note that $\bar{\mathcal{L}}^*(z) = \sum_x -q(x|z) \log p(x|z)$, the cross-entropy between the generative distribution $q_{X|Z=z}$ and the predictive distribution $p_{X|Z=z}$, and is therefore lower-bounded, for any z , by the entropy of the generative distribution, $\sum_x -q(x|z) \log q(x|z)$. For the Markov chain in particular, we obtain

$$\begin{aligned}\bar{\mathcal{L}}^*(j) &= \sum_{i=1}^N -g_{ij} \log a_{ij}, \\ \underline{\mathcal{L}}_*(i) &= \sum_{j=1}^N -g_{ij}^\dagger \log a_{ij}.\end{aligned}\tag{9}$$

By averaging over both variables using either the observer's or the generative models, we can obtain several variants of the entropy rate in addition to the standard definition. Most of these do not have any obvious interpretation, but the following are suggestive of quantities that might be relevant to an observer:

$$\begin{aligned}\bar{\mathcal{L}}_* &= \sum_{x,z} \mathcal{L}(x|z)p(x|z)q(z), \\ \bar{\mathcal{L}}_*^* &= \sum_{x,z} \mathcal{L}(x|z)q(x,z).\end{aligned}\tag{10}$$

The first of these, $\bar{\mathcal{L}}_*$ is the average level of uncertainty about the next symbol experienced by the observer while processing sequences from the generative model. The second, $\bar{\mathcal{L}}_*^*$ is the average level of surprise experienced by that observer, and is bounded from below by the entropy rate of the generative model. The two are distinct: it is possible for an observer with bad model to be very certain, but wrong, about each coming symbol and thus be continually surprised. Conversely, the observer's model could make very broad, uncertain predictions and yet always find that the most likely predicted symbol appears. In the Markov chain, these measures evaluate to

$$\begin{aligned}\bar{\mathcal{L}}_* &= \sum_{i,j} \mathcal{L}(i|j)a_{ij}\pi_j^g, \\ \bar{\mathcal{L}}_*^* &= \sum_{i,j} \mathcal{L}(i|j)g_{ij}\pi_j^g,\end{aligned}\tag{11}$$

where $\mathcal{L}(i|j) = -\log a_{ij}$ and π^g is the stationary distribution resulting from the generative transition matrix g .

4.2. Predictive information-based quantities

We can repeat the same process of taking averages with respect to the generative distributions using the predictive information $\mathcal{I}(x|z)$ instead of the surprisingness $\mathcal{L}(x|z)$. For the remainder of this section, we give the results for the Markov chain in parallel with the general expressions, with $\mathcal{I}(i|j) = \sum_k a_{ki}(\log a_{ki} - \log[a^2]_{kj})$.

In addition to the quantities already defined in (7), we obtain two variants of the symbol-specific average predictive information measures.

$$\begin{aligned}\bar{\mathcal{I}}^*(z) &= \sum_x \mathcal{I}(x|z)q(x|z), & \bar{\mathcal{I}}^*(j) &= \sum_i \mathcal{I}(i|j)g_{ij}, \\ \underline{\mathcal{I}}_*(x) &= \sum_z \mathcal{I}(x|z)q(z|x), & \underline{\mathcal{I}}_*(i) &= \sum_j \mathcal{I}(i|j)g_{ij}^\dagger.\end{aligned}\tag{12}$$

These are the average predictive information gained after z and the average informativeness of x . Of the global measures of predictive information rate, there are again two new variants that are readily interpretable:

$$\begin{aligned}\bar{\mathcal{I}}_* &= \sum_{x,z} \mathcal{I}(x|z)p(x|z)q(z), & \bar{\mathcal{I}}_* &= \sum_{i,j} \mathcal{I}(i|j)a_{ij}\pi_j^g, \\ \underline{\mathcal{I}}_*^* &= \sum_{i,j} \mathcal{I}(i|j)g_{ij}\pi_j^g, & \underline{\mathcal{I}}_*^* &= \sum_{x,z} \mathcal{I}(x|z)q(x,z).\end{aligned}\tag{13}$$

If $\bar{\mathcal{I}}(z) = \sum_x \mathcal{I}(x|z)p(x|z)$ is thought of as the ‘information expectancy’ engendered by the context $Z = z$, then $\bar{\mathcal{I}}_*$ is the average information expectancy experienced by observer while processing the data. In contrast, $\underline{\mathcal{I}}_*^*$ is the average information actually received per symbol as measured by changes in beliefs about the future of the sequence.

4.3. Effects of learning Markov chains

Using the subjective information measures defined above, we can examine how an observer’s assessment of a sequence drawn from a Markov chain changes as the observer gradually modifies its transition matrix to match that of the generative process. If the observer’s model converges to the true transition matrix, then its estimates of the entropy and predictive information rates of the process will converge to those of the generative model, but depending on the observer’s initial model and the learning process, these estimates will follow a certain trajectory. In particular, the observer’s subjective predictive information rate may increase or decrease in response to adaptive learning.

We take a Bayesian approach to learning Markov chain parameters from observations: the observer’s beliefs are represented by a distribution over possible transition matrices, and this distribution is updated using Bayes’ rule after each symbol is observed. We can allow for the possibility that the transition matrix might change over time by broadening the current distribution over transition matrices between each observation, approximating a model in which transition matrix follows a random walk. In practice, this means that the system is able to ‘forget’ about the distant past and estimate a transition matrix fitted to more recent observations.

For computational convenience we represent the observer’s beliefs about the transition matrix with product of Dirichlet distributions, one for each column of the

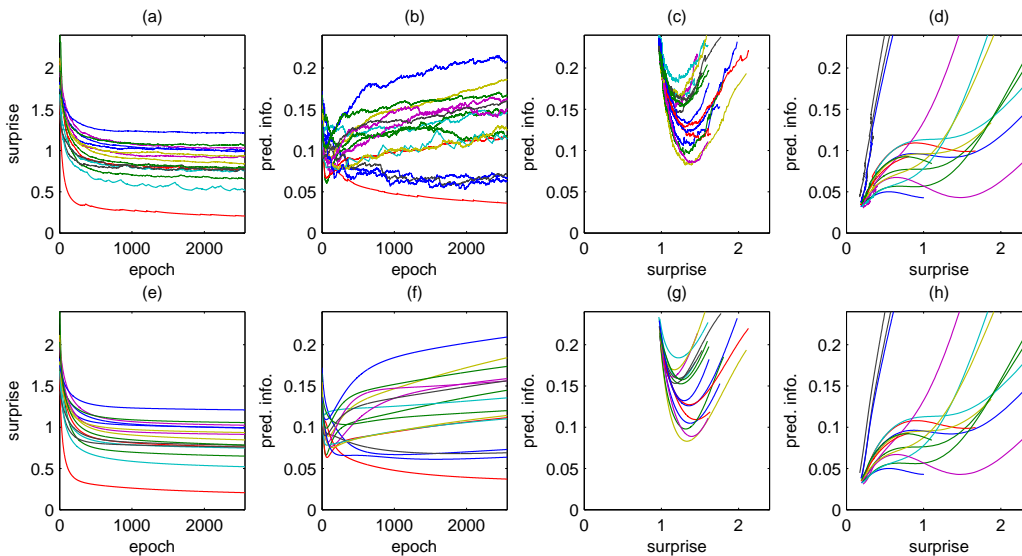


Figure 6. Learning dynamics in adaptive Markov chain system. The upper row shows the actual stochastic learning while the lower shows the idealised deterministic learning. Plots (a/b/e/f) show multiple runs starting from the same initial condition but using different generative transition matrices. Plots (c/d/g/h) show multiple runs starting from different initial conditions and converging on two transition matrices with (c/g) high or (d/h) low PIR respectively.

In (a/e) the average subjective surprisingness tends to decrease as it should since this the objective of learning. In (b/f), the subjective predictive information rate, after an initial transient phase, can go up or down depending on the generative system. The two target systems in (c/g) and (d/h) correspond to the highest and lowest lines in (a/b/e/f).

transition matrix, that is,

$$p(a|\theta) = \prod_{j=1}^N p_{\text{Dir}}(a_{:j}|\theta_{:j}), \quad (14)$$

where $a_{:j}$ is the j^{th} column of a and θ is an $N \times N$ matrix of parameters such that $\theta_{:j}$ is the parameter tuple for the N -component Dirichlet distribution p_{Dir} ,

$$p_{\text{Dir}} : \mathbb{R}^N \times \mathbb{R}^N \rightarrow \mathbb{R}, \quad p_{\text{Dir}}(\alpha|\varphi) = \frac{1}{B(\varphi)} \prod_{i=1}^N \alpha_i^{\varphi_i - 1}, \quad (15)$$

where $B : \mathbb{R}^N \rightarrow \mathbb{R}$ is the multinomial Beta function. The ‘forgetting’ step is implemented using the mapping

$$\theta_{ij} \mapsto \frac{\theta_{ij}}{1 + \beta\theta_{ij}}, \quad (16)$$

β is a parameter which controls the forgetting rate. Alternative mappings could be chosen here instead; for example, the mapping

$$\theta_{ij} \mapsto \zeta_{ij} + \frac{\theta_{ij} - \zeta_{ij}}{1 + \beta(\theta_{ij} - \zeta_{ij})}. \quad (17)$$

would cause the parameter matrix to evolve towards a ‘background’ state represented by ζ . After the ‘forgetting’ step, the next observed symbol provides fresh

evidence about the current transition matrix, which enables the observer to update its belief state. The choice of the Dirichlet distribution (being the conjugate prior of the multinomial distribution) makes these updates particularly simple: on observing the symbol i following symbol j , we increment θ_{ij} by 1. As a check, this stochastic online learning can be compared with what we would expect on average by replacing the single element increment with the mapping $\theta \mapsto \theta + g\pi^g$.

We applied the above system using many combinations of generative transition matrix g and initial observer state θ , both drawn at random. In each case, the evolution of the average surprisingness $\overline{\mathcal{L}}_*$ and the predictive information rate $\overline{\mathcal{I}}_*$ was recorded. Some of the aggregate results are shown in fig. 6.

Notably, we find that, starting from a fixed belief state and depending on the statistics of the generative system, the subjective predictive information rate can increase or decrease as learning proceeds. This sort of behaviour is suggestive of the effect mentioned in § 2.2, where human subjects change their assessment of aesthetic value after prolonged or repeated exposure to a stimulus.

5. Experiments with minimalist music

Returning to our original goal of modelling the perception of temporal structure in music, we computed dynamic information measures for two pieces of minimalist music by Philip Glass, *Two Pages* (1969) and *Gradus* (1968). Both are monophonic and isochronous, and so can be represented as a sequence of symbols where each symbol stands for one note and time maps identically onto position in the sequence. Hence, the pieces can be represented very simply yet remain ecologically valid examples of ‘real’ music.

Music in the minimalist style was specifically chosen because, more than other styles of music, minimalist music tends to be constructed around patterns that are introduced in each piece as it develops, and rather less on pre-existing conventions and stylistic norms. In terms of the present discussion, we would say that in a minimalist piece, the significant expectations arise from the apprehension of regularities observed in that piece as it develops, relying less on statistical regularities representative of a particular style, such as Baroque music or Blues, which a listener must have previously internalised to fully appreciate such styles. To put it more succinctly, the composition relies more on *intra-* as opposed to *extra-*opus stylistic norms. This means that, in our analysis, we can start with a ‘vanilla’ model, which, though capable of learning, does not initially embody any stylistic expectations such as might be gained by training on a corpus of music in a particular style.

Having said that, it should be noted that *Two Pages* embodies this ideal more than *Gradus*, the latter relying somewhat on expectations generated by familiarity with tonal music (20).

5.1. Methods

Since the aim of this experiment was not to find the best fitting model of the music, but rather to examine the behaviour of the dynamic information measures, we used the adaptive Markov chain model analysed in § 3 and § 4.3. Whilst Markov chains are not necessarily good models of music, using them does keep the computation of the various information measures relatively simple.

For *Two pages*, the distribution spreading map (16) was used with β set to 0.0004 and the Dirichlet parameters θ_{ij} initialised to 0.3 for all i, j . For *Gradus*, the pa-

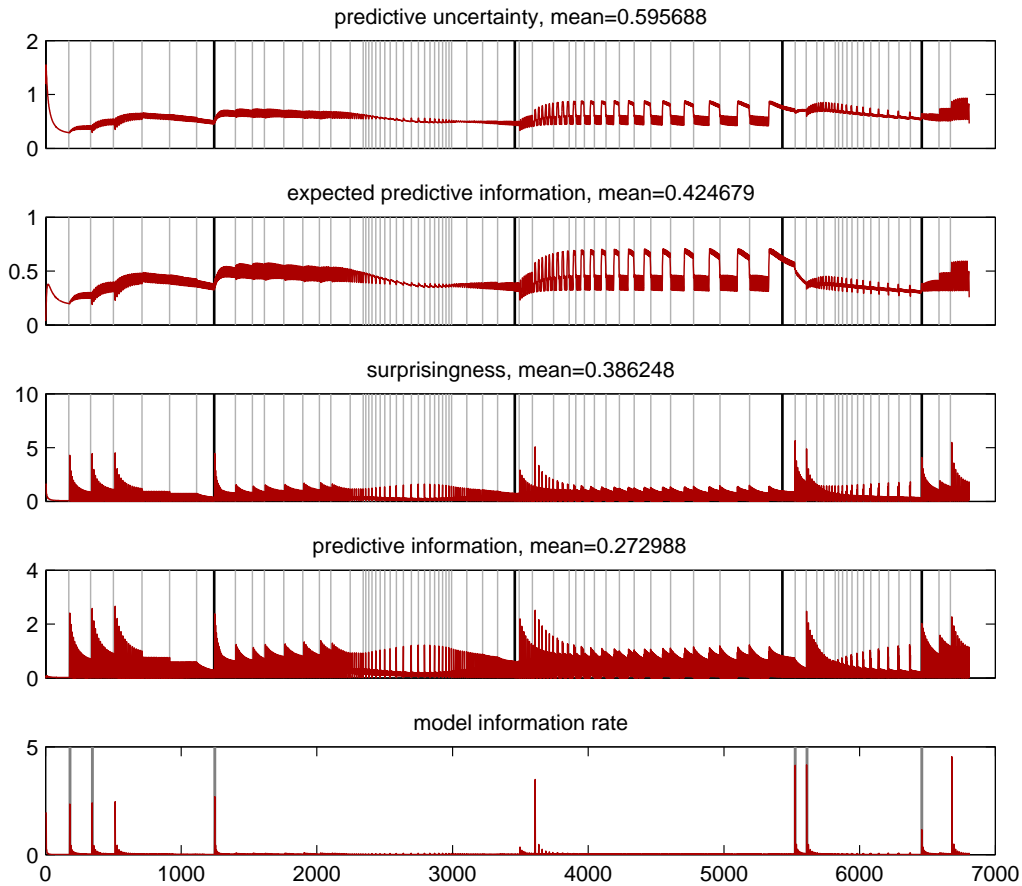
Two Pages

Figure 7. Analysis of *Two Pages*. In all panels, the thick vertical lines indicate the part boundaries as indicated in the score by the composer. The thin grey lines in the top four panels indicate changes in the melodic ‘figures’ of which the piece is constructed. In the bottom panel, the thin grey lines indicate the six most surprising moments selected by Keith Potter. All information measures are in nats.

parameters were roughly hand optimised to minimise the mean of the surprise under the constraint that all elements of θ be initialised to the same value. The mean surprisingness reached 1.29 nats/symbol with $\beta = 0.09$ and all entries of θ initialised to 0.02. By way of comparison, the multiple viewpoint variable order Markov model applied to *Gradus* in (20) achieved an average of 1.56 nats/symbol. Both of these are much lower than the $\log 12 = 2.48$ nats/symbol that would be obtained with a naïve encoding of the sequence (there are 12 symbols including one for rests), and less than the 2.29 nats/symbol that would be obtained using an encoding based on the marginal distribution of symbols.

5.2. Results

Traces of some of the dynamic information measures are shown in fig. 7 and fig. 8, along with some structural information about the pieces. In the case of *Two pages*, the correspondence between the information measures and the structure of the piece is very close. In particular, there is good agreement between the six ‘most surprising moments’ we asked music theorist and expert on minimalist music Keith Potter to choose, and the signal which tracks information gained about model’s parameters, i.e., the transition matrix. What appears to be an error in the detection of the major part boundary—between events 5000 and 6000 in fig. 7—actually raises a

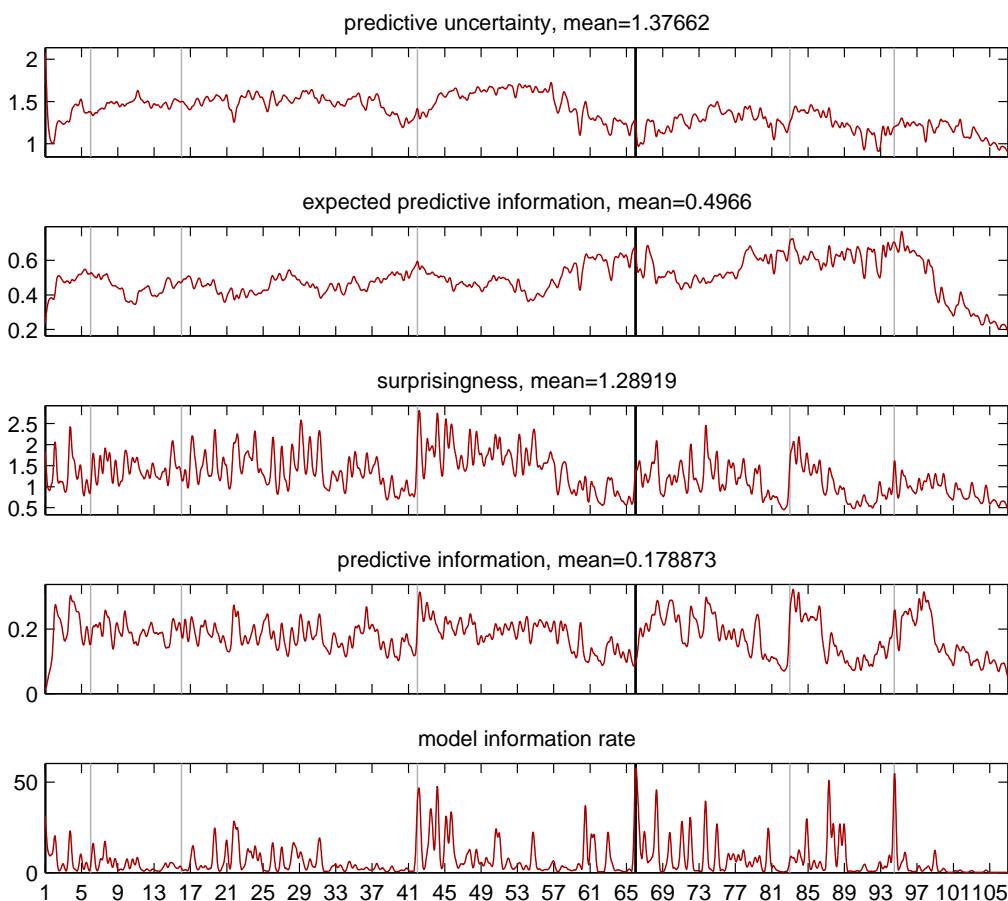
Gradus

Figure 8. Analysis of *Gradus*. In all panels, the thick black vertical lines indicate the part boundaries as indicated in the score by the composer. The thin grey lines indicate a segmentation given by Keith Potter. Note that the traces were smoothed with a Gaussian window about 12 events wide to make them more legible. All measures are in nats.

known anomaly in the score, where Glass places the boundary several events before there is any change in the pattern of notes. Alternative analyses of *Two Pages* place the boundary in agreement with peak in our surprisingness signal. There is also some noticeable structure in the predictive uncertainty and expected predictive information signals in the third section, where there is a clear alternation between two levels in each of the gradually lengthening figures. The first part of each figure consists of a pattern similar to that which opens the piece and creates the lower level of uncertainty, while the second half consists of a periodic cycling through 6 higher pitches and creates higher levels of uncertainty. In the listener, this patterns tends to create a feeling of tension, as the exact repetitions give no indication as to when the cycling will end and a new figure begin.

Gradus is much less systematically structured than *Two Pages*, and relies more on the conventions of tonal music, which are not represented the model. The information dynamic analysis shown in fig. 8 does not have such a transparent interpretations as that of *Two Pages*; nonetheless, there are many points of correspondence between the analysis and segmentation given by Keith Potter (20).

For example, peaks in the model information rate at bars 42 and 66 mark major sectional boundaries in the piece. The boundary at bar 83 is less marked in the model information signal but clearly visible in the surprisingness and predictive information signals. The major sections are visible in the broad arch-shaped

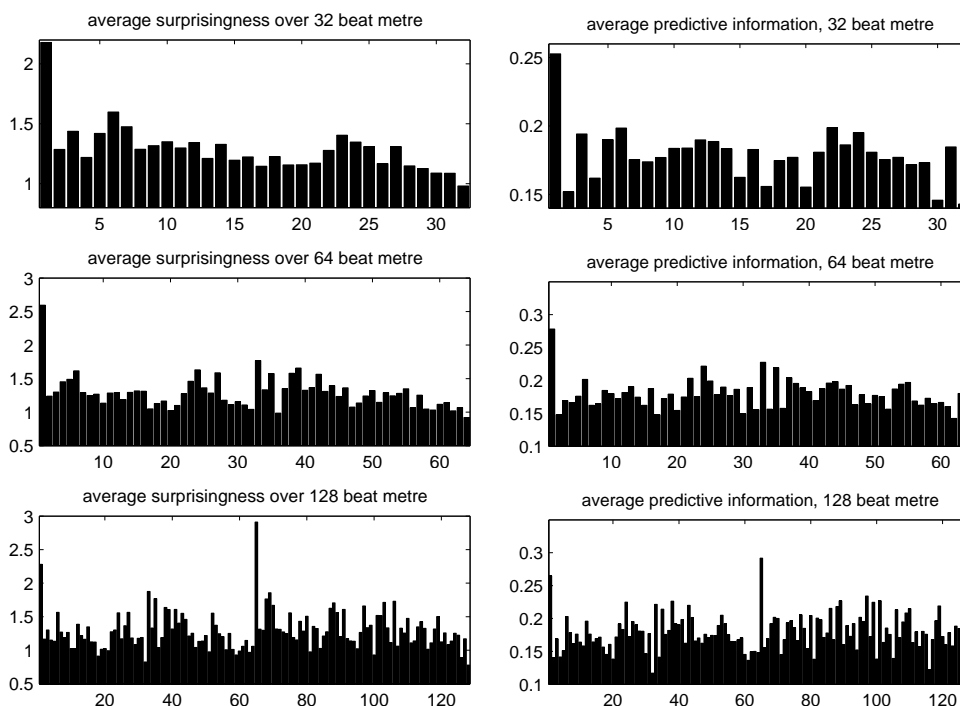


Figure 9. Information signals averaged over events at equivalent metrical positions assuming bar lengths of 32, 64, and 128 beats. (The notated bar length is 32 beats.)

developments of the predictive uncertainty signal.

Peaks at bars 4, 21, 23, 50, 60, 66, 71 and 87 all coincide with the introductions of new pitches, or in some cases, the re-introduction of a pitch that had been absent for some time. The peak at bar 44 marks the noticeable introduction of a rising sharp fourth from $C\sharp$ to G. Other peaks, such as those at bars 6, 7, 17, 19, 26 and 68 are related to changes in the melodic pattern, or in the case of the peak in the surprisingness signal around bar 35, a switch to a more fragmented rhythm with greater numbers of rests.

Towards the end of the piece, the peak at 94 marks the very prominent first occurrence of the repeated notes which bring the piece to a close, after which all the information measures begin to tail-off.

In addition to these traces of the overall structural development, there appear to be correlations between the predictive uncertainty signal and the perception of rhythmic and metrical stress. In several places where there is a strong sensation of duple time (e.g., the end of bar 7, the first half of bar 18, and in bars 72, 75 and 80), there is an accompanying pattern of alternating high and low predictive uncertainty. This is not visible in the figure because it occurs at the level of individual notes, but we observe that the stress is felt on the note *following* the note which creates higher uncertainty.

Furthermore, evidence of the 32-beat metre can be found by computing the averages of the dynamic information signals for notes at each of the 32 metrical positions. The first note of each bar is, on average (with respect to this particular model), more surprising (by approximately 0.9 nats), and more informative (by about 0.1 nats) than the other metrical positions, all of which are roughly equal according to both measures (see fig. 9). There is some evidence of a 64-beat hypermetre, but when the analysis is performed at a 128 beats, the dominant periodicity appears to remain at 64 beats. Not shown in the figure, the predictive uncertainty tends to be highest before the first note of each bar and lowest after the first note of each bar.

See (20) for further musicological analysis of *Gradus*.

6. Related work

Our definitions of the predictive information and predictive information rate are distinct from the predictive information of Bialek *et al* (21). They too consider stationary random processes, but proceed by examining the entropy of a segment of finite duration T , which, given the assumption of stationarity, will be a function of T alone, say $\mathcal{S}(T)$. This entropy will increase with increasing T , tending towards a linear growth at a rate equal to the entropy rate of the process. The mutual information between two adjacent segments, of duration T and T' respectively, can be expressed in terms of \mathcal{S} . Bialek *et al* define the predictive information as the limit of this as T' tends to infinity:

$$I_{\text{pred}}(T) = \lim_{T' \rightarrow \infty} \mathcal{S}(T) + \mathcal{S}(T') - \mathcal{S}(T + T'). \quad (18)$$

As T increases, $I_{\text{pred}}(T)$ may tend to a finite limit (which might be zero) or increase indefinitely, tending to logarithmic or fractional power-law growth. The type of growth characterises a fundamental aspect of the stochastic complexity of process. Though I_{pred} fits naturally into the information dynamics ‘toolbox’, we would argue that other measures such as the ones we describe should also be considered, since $I_{\text{pred}}(T)$ is a *global* measure which applies to the random process as a whole, not to specific realisations, much less to specific instants within a realisation.

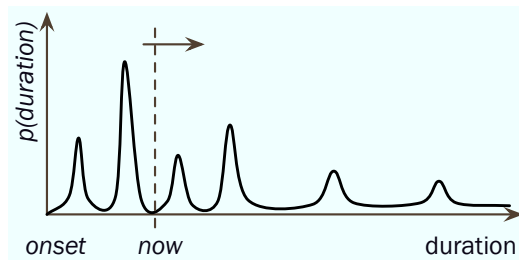
Dubnov (22) proposes an ‘information rate’ which, in our notation, can be written as $I(S_t, S_{-\infty:t-1})$, that is, the mutual information between the past and the present. For a Markov chain, this reduces to $\mathcal{H}(\pi^a) - \dot{\mathcal{H}}(a)$, where $\mathcal{H}(\pi^a)$ is the entropy of the equilibrium distribution π^a . Dubnov argues that this has the ‘inverted-U’ characteristic discussed previously in §2.4, but in the case of Markov chains at least, the effect is not what one would expect: certainly, Dubnov’s information rate is zero when each event is statistically independent of the previous one, (i.e. when the columns of the transition matrix are all identical) but the maximal information rate is reached by simultaneously minimising the entropy rate and maximising the entropy of the equilibrium distribution. This corresponds to a Markov chain where the equilibrium distribution is uniform, but after a single observation, we become able to predict the sequence reliably; a deterministic cycling through the states will have this property. One would expect such a predictable sequence to be rather uninteresting, and indeed, our PIR is zero in these cases.

The idea of measuring information gained about model parameters as the KL divergence between prior and posterior distributions is equivalent to Itti and Baldi’s ‘Bayesian surprise’ (16).

Eerola *et al* (8) propose a similar approach to ours, emphasising the need for dynamic probability models when judging uncertainty and predictability of musical patterns. They also describe experimental methods for assessing these quantities in human listeners. However, they do not explore the possibilities for multiple information measures or consider the concept of predictive information.

Levy and Jaeger (23) study spoken language using a measure of information corresponding to surprisingness (the negative log probability of a word given the previous words) and show that in certain cases speakers choose their words in order to achieve a constant information rate.

Our approach to the perception of musical structure is very much in the same spirit of that of David Huron as expounded in his book *Sweet Expectations* (24).



Huron reports and summarises many experiments showing the human subjects do indeed behave as if they have internalised the statistical regularities present in music, and also discusses how and why the perception of statistical structure (Berlyne’s collative variables again) might be closely related to affect and emotional response. For example, Huron’s suggested explanation of the *qualia* induced by different scale degrees in a tonal setting is in terms of the collative characteristics of each pitch, such as the entropy of the next pitch conditioned on this one (i.e. $\bar{\mathcal{L}}(z)$ as defined in §2), and whether or not a note tends to be surprising when it occurs (i.e. $\underline{\mathcal{L}}(x)$).

7. Discussion and conclusions

We have described an approach to the analysis of temporal structure based on an information-theoretic assessment made from the point of view an observer that updates its probabilistic model of a process dynamically as events unfold. In principle, any dynamic probabilistic model can be given this treatment. In this paper, we have examined the information dynamics of Markov chains, found an intriguing inverted-‘U’ relationship between the entropy rate and the PIR, and applied the method to the analysis of minimalist music, with some encouraging results but raising many questions and suggesting several possible developments.

Firstly, we would like to extend the analysis to more complex models such as those involving time-dependent latent variables, like HMMs. Especially relevant for music are models where events can have different durations, raising the question of how much information arrives while the observer waits for a unknown amount of time for the next event?

Secondly, since pieces of music are relatively short compared with amount of experience required to become familiar with musical styles, it will be necessary to collect models pre-trained on various style-specific corpora to act as the starting point for the processing of a particular piece.

Thirdly, to assess the cognitive relevance of our approach, we are planning experiments with human subjects to (a) search for physical correlates of the dynamic information measures, e.g. in EEG data, and (b) determine whether or not there is any relationship between the predictive information rates and the subjective experience of ‘interestingness’ and aesthetic value.

Fourthly, though we have not touched on variation in event durations, this this is likely to be an important aspect of the information dynamics of music, since rhythm is fundamental to music. The simple fact that usually we do not know how a long a note will be while it is sounding implies that we are receiving information (about the duration) even while, ostensibly, nothing is happening. The information rate will depend on the observer’s probability distribution over possible not durations, as illustrated in fig. 7.

In closing, we would like to cite some suggestive remarks from philosophers of music which have some resonance with what we are proposing. Davies (25) reviews

a range of literature on musical affect under the heading of ‘contour theories’, which is meant to convey the notion of a curve in an abstract space with time along one axis and whose shape captures some structural essence of the music. For example, Langer (26) discusses a ‘morphology of feelings’, which operates at the level of ‘patterns . . . of agreement and disagreement, preparation, fulfilment, excitation, sudden change, etc.’, arguing that these structures are relevant because they ‘exist in our minds as “amodal” forms, common to both music and feelings.’ Stern (27) used the term ‘vitality effects’ to describe ‘qualities of shape or contour, intensity, motion, and rhythm—“amodal” properties that exist in our minds as dynamic and abstract, not bound to any particular feeling or event.’ For example, ‘bursting’ could describe bursting into tears or laughter, a bursting watermelon, a burst of speed, a sforzando, and so on. Other examples include ‘surging’, ‘fading’, being ‘drawn out’ etc. Whilst such speculations are somewhat outside the scope of this paper, we do notice a common thread in the idea of an ‘amodal’ dynamic representation capturing patterns of change at an abstract level, something for which the information-dynamic approach may well provide a quantitative basis.

8. Acknowledgements

This research was supported by EPSRC grant GR/S82213/01. Thanks are also due to Keith Potter, Marcus Pearce, and Geraint Wiggins (Goldsmiths’ College, University of London) for providing the structural descriptions of *Two Pages* and *Gradus*.

References

- [1] D. E. Berlyne. *Aesthetics and Psychobiology*. Appleton Century Crofts, New York, 1971.
- [2] Leonard B. Meyer. *Music, the arts and ideas: Patterns and Predictions in Twentieth-century culture*. University of Chicago Press, 1967.
- [3] Eugene Narmour. *Beyond Schenkerism*. University of Chicago Press, 1977.
- [4] E. Hanslick. *On the musically beautiful: A contribution towards the revision of the aesthetics of music*. Hackett, Indianapolis, IN, 1854/1986.
- [5] Richard T. Cox. Probability, frequency and reasonable expectation. *American Journal of Physics*, 14:1–13, 1946.
- [6] Edwin T. Jaynes. How does the brain do plausible reasoning? In G. J. Erickson and C. R. Smith, editors, *Maximum-Entropy and Bayesian Methods in Science and Engineering*. Kluwer Academic, 1988.
- [7] J. R. Saffran, E. K. Johnson, R. N. Aslin, and E. L. Newport. Statistical learning of tone sequences by human infants and adults. *Cognition*, 70(1):27–52, 1999.
- [8] T. Eerola, P. Toiviainen, and C. L. Krumhansl. Real-time prediction of melodies: Continuous predictability judgments and dynamic models. In C. Stevens, D. Burnham, G. McPherson, E. Schubert, and J. Renwick, editors, *Proceedings of the 7th International Conference on Music Perception and Cognition (ICMPC7)*, Sydney, Australia, 2002. Causal Productions.
- [9] Daryl Conklin and Ian H. Witten. Multiple viewpoint systems for music prediction. *Journal of New Music Research*, 24(1):51–73, 1995.
- [10] D. Ponsford, G. A. Wiggins, and C. S. Mellish. Statistical learning of harmonic movement. *Journal of New Music Research*, 28(2):150–177, 1999. Also available as Research Paper 874, from the Division of Informatics, University of Edinburgh.
- [11] Marcus T. Pearce. *The Construction and Evaluation of Statistical Models of Melodic Structure in Music Perception and Composition*. PhD thesis, Department of Computing, City University, London, 2005.
- [12] Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–656, 1948.
- [13] Abraham Moles. *Information Theory and Esthetic Perception*. University of Illinois Press, 1966.
- [14] J. E. Cohen. Information theory and music. *Behavioral Science*, 7(2):137–163, 1962.
- [15] Robert E. Kass and Adrian E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430), 1995.
- [16] Laurent Itti and Pierre Baldi. Bayesian surprise attracts human attention. In *Advances Neural Information Processing Systems (NIPS 2005)*, 2005.
- [17] W. Wundt. *Outlines of Psychology*. Englemann, Leipzig, 1897.
- [18] Leonard B. Meyer. Music and emotion: Distinctions and uncertainties. In Juslin and Sloboda (28), chapter 15, pages 341–360.

- [19] Bruno de Finetti. *Theory of Probability*. John Wiley and Sons, New York, 1975.
- [20] Keith Potter, Geraint A. Wiggins, and Marcus T. Pearce. Towards greater objectivity in music theory: Information-dynamic analysis of minimalist music. *Musicae Scientiae*, 11(2):295–324, 2007.
- [21] William Bialek, Ilya Nemenman, and Naftali Tishby. Predictability, complexity, and learning. *Neural Computation*, 13:2409–2463, 2001.
- [22] Shlomo Dubnov. Spectral anticipations. *Computer Music Journal*, 30(2):63–83, 2006.
- [23] Roger Levy and T. Florian Jaeger. Speakers optimize information density through syntactic reduction. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*. MIT Press, Cambridge, MA, 2007.
- [24] David Huron. *Sweet Expectations*. MIT Press, 2006.
- [25] Stephen Davies. Philosophical perspectives on music’s expressiveness. In Juslin and Sloboda (28), chapter 2, pages 23–44.
- [26] Susanne K. Langer. *Philosophy in a new key*. Harvard University Press, Cambridge, MA, 1957.
- [27] D. Stern. *The Interpersonal World of the Infant*. Academic Press, London, 1985.
- [28] Patrick N. Juslin and John A. Sloboda, editors. *Music and Emotion — Theory and Research*. Oxford University Press, 2004.

Appendix A. Derivations for Markov Chains

Let $S : \Omega \rightarrow \mathcal{A}^\infty$ be a random process whose realisations are infinite sequences of elements taken from an alphabet \mathcal{A} . The t^{th} element of the sequence is represented by the random variable $S_t : \Omega \rightarrow \mathcal{A}$. A realisation of the random process is a sequence $s = S(\omega) \in \mathcal{A}^\infty$, where ω is a sample drawn from a probability space on Ω . We assume that \mathcal{A} contains N elements $\{\sigma_1, \dots, \sigma_N\}$. If S is a Markov chain, then the process can be parameterised in terms of a transition matrix $a \in \mathbb{R}^{N \times N}$ and an initial distribution $b \in \mathbb{R}^N$ for the first element of the sequence:

$$a_{ij} \triangleq \Pr(S_{t+1} = \sigma_i | S_t = \sigma_j), \quad (\text{A1})$$

$$b_i \triangleq \Pr(S_1 = \sigma_i). \quad (\text{A2})$$

Note that $\Pr(\psi)$ denotes the probability that ψ is true where ψ is logical formula. Similarly, $\Pr(\psi|\phi)$ denotes the probability of ψ conditioned on the truth of ϕ . Probability distribution functions will be written as a p with a subscript to indicate from which random variables the arguments are intended to be drawn, e.g., $p_{S_t} : \mathcal{A} \rightarrow \mathbb{R}$ is the marginal distribution function of the t^{th} element of the chain and thus $p_{S_t}(\sigma_i)$ is the probability that S_t takes the value σ_i .

The equilibrium or stationary distribution $\pi^a \in \mathbb{R}^N$ of the Markov chain is defined by the condition $a\pi^a = \pi^a$, which implies that π^a is an eigenvector of the transition matrix a with eigenvalue 1. In order that the equilibrium distribution be unique, we require that the Markov chain be *irreducible*, i.e., that every state is potentially reachable from every other state.

Since we want the Markov chain to be stationary and to have a well defined equilibrium distribution π^a , we must have $\Pr(S_t = \sigma_i) = \pi_i^a$ for all t including $t = 1$. Hence, $b = \pi^a$, which is in turn a function of a .

A.1. Entropy and entropy rate

Having found the equilibrium distribution, we can derive the entropy of any single element S_t of the chain taken in isolation. For any t , $H(S_t) = \mathcal{H}(\pi^a)$, where \mathcal{H} is the Shannon entropy function defined as

$$\mathcal{H} : \mathbb{R}^N \rightarrow \mathbb{R}, \quad \mathcal{H}(\theta) = \sum_{i=1}^N -\theta_i \log \theta_i, \quad (\text{A3})$$

The conditional entropy $H(S_{t+1}|S_t)$ can be derived by considering the joint distribution of S_{t+1} and S_t :

$$H(S_{t+1}|S_t) = \sum_{i=1}^N \sum_{j=1}^N -\Pr(S_{t+1}=\sigma_i \wedge S_t=\sigma_j) \log \Pr(S_{t+1}=\sigma_i | S_t=\sigma_j) \quad (\text{A4})$$

Hence, we can define a function $\dot{\mathcal{H}} : \mathbb{R}^{N \times N} \rightarrow \mathbb{R}$ such that $H(S_{t+1}|S_t) = \dot{\mathcal{H}}(a)$:

$$\dot{\mathcal{H}}(a) \triangleq \sum_{i=1}^N \sum_{j=1}^N -a_{ij} \pi_j^a \log a_{ij} \quad (\text{A5})$$

This is independent of t and for a first order Markov chain yields the entropy rate of the process.

A.2. Predictive information rates

The average predictive information rate (APIR), using X , Y , and Z to stand for the present, future and past respectively, can be written in several ways including

$$\begin{aligned} I(X, Y|Z) &= H(Y|Z) - H(Y|X, Z) \\ &= H(X|Z) - H(X|Y, Z). \end{aligned} \quad (\text{A6})$$

At this point we will assume without loss of generality that the Markov chain extends infinitely in both directions and that the current time is zero, so that $Z = S_{-\infty:-1}$, $X = S_0$, and $Y = S_{1:\infty}$.

Now, in general, if three variables A , B and C are such that A and C are conditionally independent given B , that is, in the commonly understood abuse of notation, $p(c|b, a) = p(c|b)$, then $H(C|B, A) = H(C|B)$:

$$\begin{aligned} H(C|B, A) &= \sum_{a,b,c} p(c|b, a)p(b, a) \log p(c|b, a) \\ &= \sum_{b,c} p(c|b) \left(\sum_a p(b, a) \right) \log p(c|b) \\ &= \sum_{b,c} p(c|b)p(b) \log p(c|b) \\ &= H(C|B). \end{aligned} \quad (\text{A7})$$

For the Markov chain, this implies that the APIR can be written as

$$\begin{aligned} I(S_0, S_{1:\infty}|S_{-\infty:-1}) &= H(S_{1:\infty}|S_{-\infty:-1}) - H(S_{1:\infty}|S_0, S_{-\infty:-1}) \\ &= H(S_{2:\infty}|S_1) + H(S_1|S_{-1}) - (H(S_{2:\infty}|S_1) + H(S_1|S_0)) \\ &= H(S_1|S_{-1}) - H(S_1|S_0) \end{aligned} \quad (\text{A8})$$

The second term is the entropy rate $\dot{\mathcal{H}}(a)$ of the Markov chain, while the first term can be identified as the entropy rate of the Markov chain obtained by taking every second element of the original chain. The transition matrix of this derived two-step chain is simply the matrix square of the original transition matrix, i.e. a^2 . If it too is irreducible, then it will have the same equilibrium distribution as the original and the average predictive information rate will be $\dot{\mathcal{H}}(a^2) - \dot{\mathcal{H}}(a)$.

The instantaneous predictive information rate for the Markov chain is derived by considering the information in the observation $S_0 = s_0$ about the entire tail of the sequence $S_{1:\infty}$ given the preceding context $S_{-\infty:-1} = s_{-\infty:-1}$. We will write this as $I(S_0 = s_0, S_{1:\infty}|S_{-\infty:-1} = s_{-\infty:-1})$ and compute it as the KL divergence between the prior $p_{S_{1:\infty}|S_{-\infty:-1} = s_{-\infty:-1}}$ and the posterior $p_{S_{1:\infty}|S_0 = s_0, S_{-\infty:-1} = s_{-\infty:-1}}$. Because of the Markov dependency structure this can immediately simplified to

$$I(S_0 = s_0, S_{1:\infty}|S_{-\infty:-1} = s_{-\infty:-1}) = D(p_{S_{1:\infty}|S_0 = s_0} || p_{S_{1:\infty}|S_{-1} = s_{-1}}). \quad (\text{A9})$$

Expanding this using the definition of the KL divergence (and dropping the subscripts of the distribution functions where the relevant random variables are clear

from the arguments) yields

$$\begin{aligned}
& D(p_{S_{1:\infty}|S_0=s_0} || p_{S_{1:\infty}|S_{-1}=s_{-1}}) \\
&= \sum_{s_{1:\infty} \in \mathcal{A}^\infty} p(s_{1:\infty}|s_0) \log \frac{p(s_{1:\infty}|s_0)}{p(s_{1:\infty}|s_{-1})} \\
&= \sum_{s_{1:\infty} \in \mathcal{A}^\infty} p(s_{2:\infty}|s_1)p(s_1|s_0) \log \frac{p(s_{2:\infty}|s_1)p(s_1|s_0)}{\sum_{s'_0 \in \mathcal{A}} p(s_{2:\infty}|s_1)p(s_1|s'_0)p(s'_0|s_{-1})} \quad (\text{A10}) \\
&= \sum_{s_1 \in \mathcal{A}} \left(\sum_{s_{2:\infty} \in \mathcal{A}^\infty} p(s_{2:\infty}|s_1) \right) p(s_1|s_0) \log \frac{p(s_1|s_0)}{\sum_{s'_0 \in \mathcal{A}} p(s_1|s'_0)p(s'_0|s_{-1})} \\
&= \sum_{s_1 \in \mathcal{A}} p(s_1|s_0) \log \frac{p(s_1|s_0)}{\sum_{s'_0 \in \mathcal{A}} p(s_1|s'_0)p(s'_0|s_{-1})}
\end{aligned}$$

This shows that the information in $S_0=s_0$ about the entire future is accounted for by information it contains about the next element of the chain. Rewritten in terms of the transition matrix, the predictive information is a function of the current and previous states alone:

$$\begin{aligned}
\mathcal{I}(i|j) &= I(S_0=\sigma_i, S_1|S_{-1}=\sigma_j) \\
&= \sum_{k=1}^N a_{ki} \log \frac{a_{ki}}{[a^2]_{kj}}. \quad (\text{A11})
\end{aligned}$$