

# INFORMATION DYNAMICS

Samer Abdallah and Mark Plumbley

Centre for Digital Music,  
Queen Mary, University of London  
Technical Report C4DM-TR07-01  
Version 1.0 – July 18, 2007

## Abstract

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.



# Information Dynamics

Samer Abdallah (samer.abdallah@elec.qmul.ac.uk)

Mark Plumbley (mark.plumbley@elec.qmul.ac.uk)

July 18, 2007

## Abstract

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.

## 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.

The business of making predictions and assessing surprise is essentially one of reasoning with degrees of belief and best quantified in terms of Bayesian probability theory [5, 6]. 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, 10, 11].

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. [13, 2, 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.

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 these measures, we obtain a representation which captures much of the significant structure of the music. 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.

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, etc. 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.

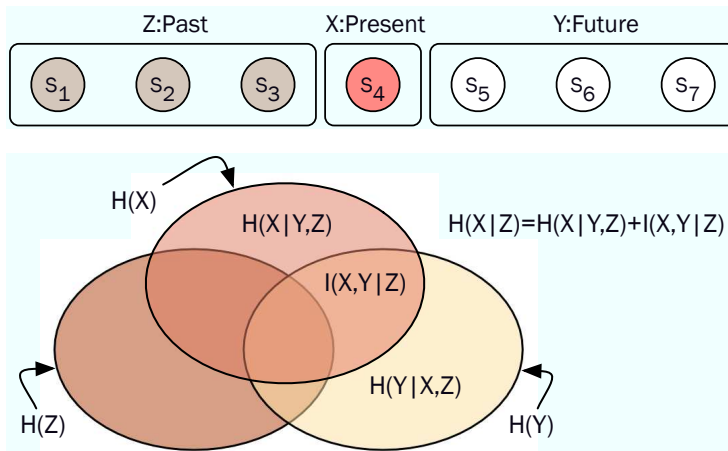


Figure 1: By grouping the elements of a random sequence into a *past*, *present*, and *future*, we can consider a number of information measures, some of which are well known, like the entropy rate  $H(X|Z)$ , and some of which have not, to our knowledge, been investigated before, such as the average 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.

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 are in a position to relate our approach with previous work in the same area. In §5 we apply the Markov chain model to minimalist music by Philip Glass and show how the information-dynamic approach yields a plausible structural analysis that largely agrees of a human expert listener. We wrap-up with discussion and conclusions in §6.

## 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, 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 *instantaneous predictive information rate* (IPIR). 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 [15], as it is a mechanism which can operate *before* the surprise occurs.

The average of the IPIR 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 *average predictive information rate* (APIR), 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  $\Theta$ . 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’.

### 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 aesthetic value, reveal an inverted ‘U’ shaped curve 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 [16]. 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 [17] hints at the same thing 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,

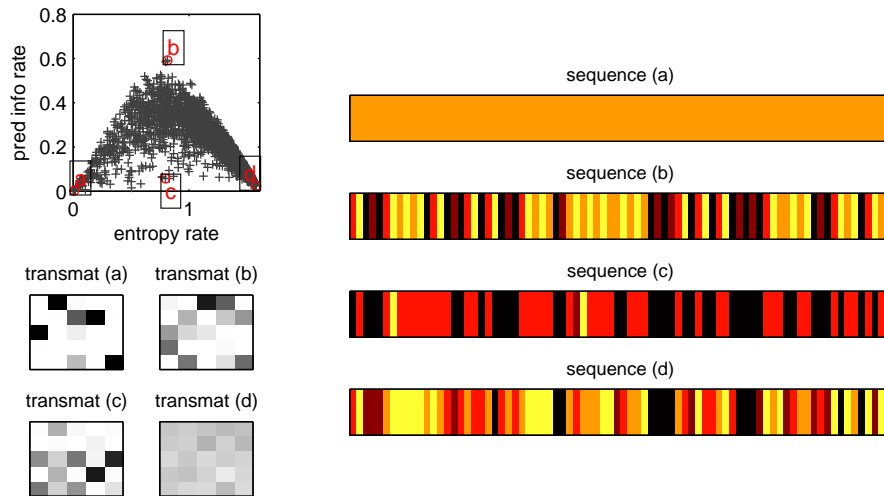


Figure 2: The space of transition matrices explored by generating them at random and plotting entropy rate vs APIR. (Note inverted ‘U’ relationship). Four of the transition matrices are shown along with sample sequences. Sequence (a) is simply the endless repetition of state 4 due to the values in the fourth column of matrix (a). Matrix (d) is almost uniform. Matrix (b) has the highest APIR.

the predictive information rate (3) seems to incorporate this balance automatically (see fig. 2), 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 APIR, this is something that could be tested experimentally with human subjects.

### 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  $t^{\text{th}}$  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  $\pi^a$  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 as a function of  $a$  alone:

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

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

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

The second four, predictive-information-based measures are

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

**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 APIR have intermediate entropy. The scatter plot in fig. 2 was obtained by generating transition matrices at random by drawing each column independently from a Dirichlet distribution. We also investigated optimising the APIR directly using a general purpose optimiser. We found that, for a range of different  $N$ , relatively sparse transition matrices maximise the APIR (see fig. 3).

## 4 Related work

Our definitions of the average and instantaneous predictive information rates are distinct from the predictive information of Bialek *et al* [18]. 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  $S(T)$ . 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

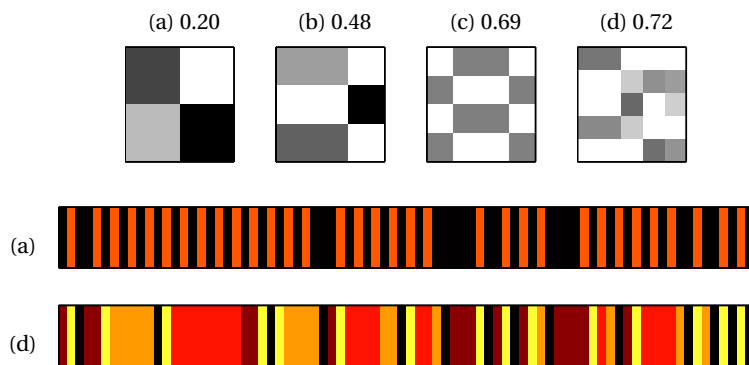


Figure 3: The results of direct numerical optimisation of the APIR for different state space sizes  $N$ . The number over each transition matrix is its APIR in nats, while the two sequences below are samples from transition matrices (a) and (d) respectively.

expressed in terms of  $\mathcal{S}$ . Bialek *et al* define the predictive information as the limit of this as  $T'$  tends to infinity:

$$I_{\text{pred}}(T) = \lim_{T' \rightarrow \infty} \mathcal{S}(T) + \mathcal{S}(T') - \mathcal{S}(T + T'). \quad (8)$$

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  $I_{\text{pred}}$  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 is applies to the random process as a whole, not to specific realisations, much less to specific instants within a realisation.

Dubnov [19] 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  $\mathcal{H}(\pi^a) - \mathcal{H}(a)$ , where  $\mathcal{H}(\pi^a)$  is the entropy of the equilibrium distribution  $\pi^a$ . 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 transtion 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 APIR 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’ [15].

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 [20] 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.

## 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.

### 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 a simple elaboration of the Markov chain analysed in §3. Whilst Markov chains are not particularly good models of music, using them does enable us to compute the various information measures in closed-form without having to make approximations.

The one elaboration on the basic Markov chain is that we allow the transition matrix to vary slowly with time in order to better fit and subsequently track changes in the sequence structure. This means that the observer’s belief state must include a probability distribution over transition matrices, which we represent as a product of Dirichlet distributions, one for each column of the transition matrix, that is,

$$p(a|\theta) = \prod_{j=1}^N p_{\text{Dir}}(a_{:,j}|\theta_{:,j}), \quad (9)$$

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

$$p_{\text{Dir}} : \mathbb{R}^N \times \mathbb{R}^N \rightarrow \mathbb{R}, \quad p_{\text{Dir}}(\alpha|\varphi) = \frac{1}{B(\varphi)} \prod_{i=1}^N \alpha_i^{\varphi_i - 1}, \quad (10)$$

where  $B : \mathbb{R}^N \rightarrow \mathbb{R}$  is the multinomial Beta function.

At each time step, the distribution over transition matrices evolves into a slightly broader one under the mapping

$$\theta_{ij} \mapsto \frac{\beta \theta_{ij}}{(\beta + \theta_{ij})}. \quad (11)$$

## Two Pages

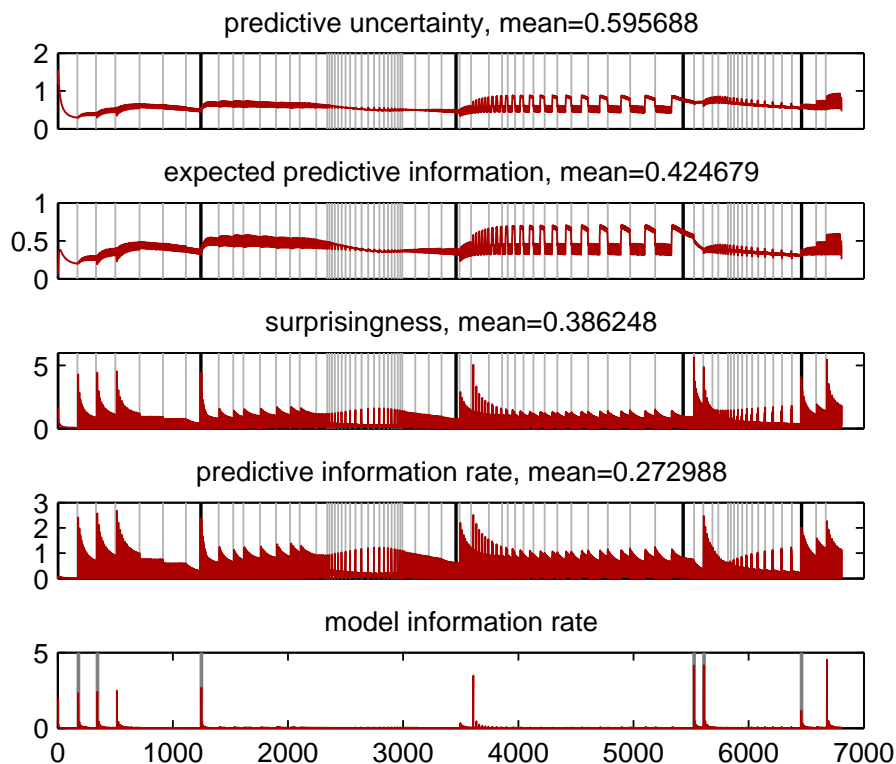


Figure 4: Analyses of *Two Pages*. In all panels, the thick vertical lines indicate the part boundaries as indicated in the score by the composer. The thin grey lines in the top four panels indicate changes in the melodic ‘figures’ of which the piece is constructed. In the bottom panel, the thin grey lines indicate the six most surprising moments selected by Keith Potter. All information measures are in nats.

This models the possibility that the transition matrix has changed;  $\beta$  is a ‘persistence’ parameter and was set to 2500 in our experiments. 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.

## 5.2 Results

Traces of some of the dynamic information measures are shown in fig. 4 and fig. 5, 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 quite 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 pa-

## Gradus

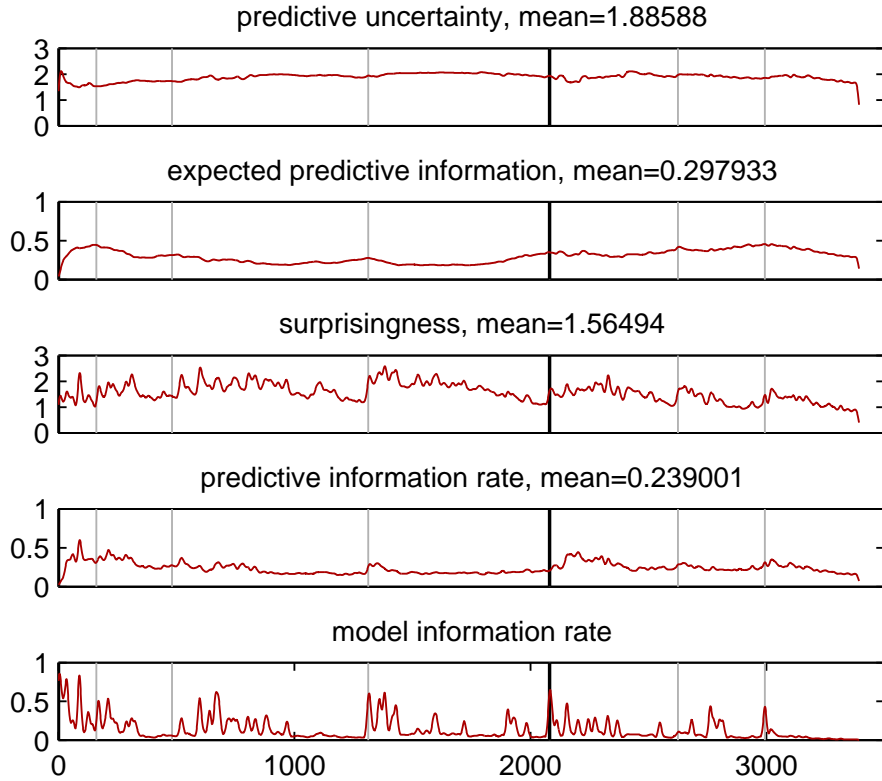


Figure 5: Analyses *Gradus*. In all panels, the thick black vertical lines indicate the part boundaries as indicated in the score by the composer. The thin grey lines indicate a segmentation given by Keith Potter. Note that the traces were smoothed with a Gaussian window about 16 events wide to make them more legible. All measures are in nats.

rameters, 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. 4—actually raises a 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.

*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. 5 does not give such a clear picture of the structure; nonetheless, there are some points of correspondence between the analysis and segmentation given by Keith Potter.

## 6 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 up-

dates 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 APIR, 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.

In closing, we would like to cite some suggestive remarks from philosophers of music which have some resonance with what we are proposing. Davies [21] 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 whos shape captures some structural essence of the music. For example, Langer [22] 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 [23] 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. Others 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.

**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] Laurent Itti and Pierre Baldi. Bayesian surprise attracts human attention. In *Advances Neural Information Processing Systems (NIPS 2005)*, 2005.

- [16] W. Wundt. *Outlines of Psychology*. Englemann, Leipzig, 1897.
- [17] Leonard B. Meyer. Music and emotion: Distinctions and uncertainties. In Juslin and Sloboda [24], chapter 15, pages 341–360.
- [18] William Bialek, Ilya Nemenman, and Naftali Tishby. Predictability, complexity, and learning. *Neural Computation*, 13:2409—2463, 2001.
- [19] Shlomo Dubnov. Spectral anticipations. *Computer Music Journal*, 30(2):63–83, 2006.
- [20] 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.
- [21] Stephen Davies. Philosophical perspectives on music’s expressiveness. In Juslin and Sloboda [24], chapter 2, pages 23–44.
- [22] Susanne K. Langer. *Philosophy in a new key*. Harvard University Press, Cambridge, MA, 1957.
- [23] D. Stern. *The Interpersonal World of the Infant*. Academic Press, London, 1985.
- [24] Patrick N. Juslin and John A. Sloboda, editors. *Music and Emotion — Theory and Research*. Oxford University Press, 2004.

## 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  $t^{\text{th}}$  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  $s = S(\omega) \in \mathcal{A}^\infty$ . 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:

$$a_{ij} \triangleq \Pr(S_{t+1} = \sigma_i | S_t = \sigma_j), \quad (12)$$

$$b_i \triangleq \Pr(S_1 = \sigma_i). \quad (13)$$

Note that  $\Pr(\psi)$  denotes the probability that  $\psi$  is true where  $\psi$  is logical formula. Similarly,  $\Pr(\psi|\phi)$  denotes the probability of  $\psi$  conditioned on the truth of  $\phi$ . 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  $t^{\text{th}}$  element of the chain and thus  $p_{S_t}(\sigma_i)$  is the probability that  $S_t$  takes the value  $\sigma_i$ .

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  $\pi^a$  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  $\pi^a$ , 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

$$\mathcal{H} : \mathbb{R}^N \rightarrow \mathbb{R}, \quad \mathcal{H}(\theta) = \sum_{i=1}^N -\theta_i \log \theta_i, \quad (14)$$

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

$$H(S_{t+1}|S_t) = \sum_{i=1}^N \sum_{j=1}^N -\Pr(S_{t+1} = \sigma_i \wedge S_t = \sigma_j) \log \Pr(S_{t+1} = \sigma_i | S_t = \sigma_j) \quad (15)$$

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

$$\dot{\mathcal{H}}(a) \triangleq \sum_{i=1}^N \sum_{j=1}^N -a_{ij} \pi_j^a \log a_{ij} \quad (16)$$

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

$$\begin{aligned} I(X, Y|Z) &= H(Y|Z) - H(Y|X, Z) \\ &= H(X|Z) - H(X|Y, Z). \end{aligned} \quad (17)$$

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  $H(C|B, A) = H(C|B)$ :

$$\begin{aligned} H(C|B, A) &= \sum_{a,b,c} p(c|b, a) p(b, a) \log p(c|b, a) \\ &= \sum_{b,c} p(c|b) \left( \sum_a p(b, a) \right) \log p(c|b) \\ &= \sum_{b,c} p(c|b) p(b) \log p(c|b) \\ &= H(C|B). \end{aligned} \quad (18)$$

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

$$\begin{aligned} I(S_0, S_{1:\infty}|S_{-\infty:-1}) &= H(S_{1:\infty}|S_{-\infty:-1}) - H(S_{1:\infty}|S_0, S_{-\infty:-1}) \\ &= H(S_{2:\infty}|S_1) + H(S_1|S_{-1}) - (H(S_{2:\infty}|S_1) + H(S_1|S_0)) \\ &= H(S_1|S_{-1}) - H(S_1|S_0) \end{aligned} \quad (19)$$

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 be simplified to

$$I(S_0 = s_0, S_{1:\infty}|S_{-\infty:-1} = s_{-\infty:-1}) = D(p_{S_{1:\infty}|S_0=s_0} \| p_{S_{1:\infty}|S_{-1}=s_{-1}}). \quad (20)$$

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

$$\begin{aligned}
& D(p_{S_{1:\infty}|S_0=s_0} || p_{S_{1:\infty}|S_{-1}=s_{-1}}) \\
&= \sum_{s_{1:\infty} \in \mathcal{A}^\infty} p(s_{1:\infty}|s_0) \log \frac{p(s_{1:\infty}|s_0)}{p(s_{1:\infty}|s_{-1})} \\
&= \sum_{s_{1:\infty} \in \mathcal{A}^\infty} p(s_{2:\infty}|s_1) p(s_1|s_0) \log \frac{p(s_{2:\infty}|s_1) p(s_1|s_0)}{\sum_{s'_0 \in \mathcal{A}} p(s_{2:\infty}|s_1) p(s_1|s'_0) p(s'_0|s_{-1})} \quad (21) \\
&= \sum_{s_1 \in \mathcal{A}} \left( \sum_{s_{2:\infty} \in \mathcal{A}^\infty} p(s_{2:\infty}|s_1) \right) p(s_1|s_0) \log \frac{p(s_1|s_0)}{\sum_{s'_0 \in \mathcal{A}} p(s_1|s'_0) p(s'_0|s_{-1})} \\
&= \sum_{s_1 \in \mathcal{A}} p(s_1|s_0) \log \frac{p(s_1|s_0)}{\sum_{s'_0 \in \mathcal{A}} p(s_1|s'_0) p(s'_0|s_{-1})}
\end{aligned}$$

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:

$$\begin{aligned}
\mathcal{I}(i|j) &= I(S_0 = \sigma_i, S_1 | S_{-1} = \sigma_j) \\
&= \sum_{k=1}^N a_{ki} \log \frac{a_{ki}}{[a^2]_{kj}}. \quad (22)
\end{aligned}$$