Information Dynamics and Temporal Structure in Music

Samer Abdallah and Mark Plumbley

Centre for Digital Music
Queen Mary, University of London
www.elec.qmul.ac.uk/digitalmusic/

December 7, 2007

Outline

Outline

Expectation and suprise in music

Music creates *expectations* of what is to come next, which may be fulfilled immediately, after some delay, or not at all. Suggested by music theorists, e.g. L. B. Meyer [Mey67] and Narmour [Nar77] but also noted much earlier by Hanslick [Han86] in the 1850s:

> *'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 . . . 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.'*

‘Unfoldingness’

Music is experienced as a phenomenon that

‘Unfoldingness’

Music is experienced as a phenomenon that ‘unfolds’

‘Unfoldingness’

Music is experienced as a phenomenon that ‘unfolds’ in

‘Unfoldingness’

Music is experienced as a phenomenon that ‘unfolds’ in time,

‘Unfoldingness’

Music is experienced as a phenomenon that ‘unfolds’ in time, rather than being apprehended as a static object presented in its entirety.

‘Unfoldingness’

Music is experienced as a phenomenon that ‘unfolds’ in time, rather than being apprehended as a static object presented in its entirety.

Meyer [Mey67] 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.

Probabilistic reasoning

Making predictions and assessing surprise is essentially reasoning with degrees of belief and (arguably) the best way to do this is using Bayesian probability theory [Cox46, Jay88].

Probabilistic reasoning

Making predictions and assessing surprise is essentially reasoning with degrees of belief and (arguably) the best way to do this is using Bayesian probability theory [Cox46, Jay88].

We suppose that familiarity with different styles of music takes the form of various probabilistic models, and that these models are adapted through listening.

Probabilistic reasoning

Making predictions and assessing surprise is essentially reasoning with degrees of belief and (arguably) the best way to do this is using Bayesian probability theory [Cox46, Jay88].

We suppose that familiarity with different styles of music takes the form of various probabilistic models, and that these models are adapted through listening.

Experimental evidence that humans are able to internalise statistical knowledge about musical: [SJAN99, ETK02]; and also that statistical models are effective for computational analysis of music, e.g. [CW95, Pea05].

Music and information theory

With probabilistic models in hand we can apply quantitative information theory: we can compute entropies, relative entropies, mutual information, and all that.

Music and information theory

With probabilistic models in hand we can apply quantitative information theory: we can compute entropies, relative entropies, mutual information, and all that.

Lots of interest in application of information theory to perception, music and aesthetics since the 50s, e.g. Moles [Mol66], Meyer [Mey67], Cohen [Coh62], Berlyne [Ber71]. (See also Bense, Hiller)

Music and information theory

With probabilistic models in hand we can apply quantitative information theory: we can compute entropies, relative entropies, mutual information, and all that.

Lots of interest in application of information theory to perception, music and aesthetics since the 50s, e.g. Moles [Mol66], Meyer [Mey67], Cohen [Coh62], Berlyne [Ber71]. (See also Bense, Hiller)

Idea is that subjective qualities and states like uncertainty, surprise, complexity, tension, and interestingness are determined by information-theoretic quantities.

Music and information theory

With probabilistic models in hand we can apply quantitative information theory: we can compute entropies, relative entropies, mutual information, and all that.

Lots of interest in application of information theory to perception, music and aesthetics since the 50s, e.g. Moles [Mol66], Meyer [Mey67], Cohen [Coh62], Berlyne [Ber71]. (See also Bense, Hiller)

Idea is that subjective qualities and states like uncertainty, surprise, complexity, tension, and interestingness are determined by information-theoretic quantities.

Berlyne [Ber71] called such quantities ‘collative variables’, since they are to do with patterns of occurrence rather than medium-specific details. *Information aesthetics*.

Probabilistic model-based observer hypothesis

- As we listen, we maintain a probabilistic model that enables us to make predictions. As events unfold, we revise our probabilistic 'belief state', including predictions about the future.

Probabilistic model-based observer hypothesis

- As we listen, we maintain a probabilistic model that enables us to make predictions. As events unfold, we revise our probabilistic 'belief state', including predictions about the future.
- Probability distributions and changes in distributions are characterised in terms of information theoretic-measures such as entropy and relative entropy (KL divergence).

Probabilistic model-based observer hypothesis

- As we listen, we maintain a probabilistic model that enables us to make predictions. As events unfold, we revise our probabilistic 'belief state', including predictions about the future.
- Probability distributions and changes in distributions are characterised in terms of information theoretic-measures such as entropy and relative entropy (KL divergence).
- The dynamic evolution of these information measures captures significant structure, e.g. events that are surprising, informative, explanatory etc.

Features of information dynamics

Abstraction: sensitive mainly to *patterns* of occurence, rather than details of which specific things occur or the sensory medium.

Features of information dynamics

Abstraction: sensitive mainly to *patterns* of occurence, rather than details of which specific things occur or the sensory medium.

Generality: applicable in principle to any probabilistic model, in particular, models with time-dependent latent variables such as HMMs. Many important musical concepts like key, harmony, and beat are essentially ‘hidden variables’.

Features of information dynamics

Abstraction: sensitive mainly to *patterns* of occurence, rather than details of which specific things occur or the sensory medium.

Generality: applicable in principle to any probabilistic model, in particular, models with time-dependent latent variables such as HMMs. Many important musical concepts like key, harmony, and beat are essentially ‘hidden variables’.

Richness: when applied to models with latent variables, can result in many-layered analysis, capturing information flow about harmony, tempo, etc.

Features of information dynamics

Abstraction: sensitive mainly to *patterns* of occurence, rather than details of which specific things occur or the sensory medium.

Generality: applicable in principle to any probabilistic model, in particular, models with time-dependent latent variables such as HMMs. Many important musical concepts like key, harmony, and beat are essentially 'hidden variables'.

Richness: when applied to models with latent variables, can result in many-layered analysis, capturing information flow about harmony, tempo, etc.

Subjectivity: all probabilities are *subjective* probabilities relative to *observer's* model, which can depend on observer's capabilities and prior experience.

Contour theories

Davies [Dav04] reviews literature on musical affect under the heading of ‘contour theories’. ‘Contour’ is a curve in an abstract space with time along one axis.

Langer [Lan57] discusses a ‘morphology of feelings’: ‘patterns . . . of agreement and disgreement, 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’s [Ste85] ‘vitality effects’: ‘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.’

Common idea of an ‘amodal’ dynamic representation capturing patterns of change at an abstract level.

Outline

Information theory primer · Entropy

Entropy is a measure of *uncertainty*. If observer expects to see x with probability $p(x)$, then

$$H(X) = \mathrm{E} - \log p(x) = \int_{\mathcal{X}} -p(x) \log p(x) \, \mathrm{d}x$$

Consider $-\log p(x)$ as the 'surprisingness' of x, then the entropy is the 'expected surprisingness'. High for spread out distributions and low for concentrated ones.

Information theory primer · Relative entropy

Relative entropy or KL divergence quantifies difference between probability distributions. If data $\mathcal{D}$ arrives, divergence between prior and posterior distributions is the amount of information in $\mathcal{D}$ *about* X:

$$D(p_{X|\mathcal{D}}||p_X) = \int_{\mathcal{X}} p_{X|\mathcal{D}}(x) \log \frac{p_{X|\mathcal{D}}(x)}{p_X(x)} \, \mathrm{d}x$$

If observing $\mathcal{D}$ causes a large change in belief about X, then $\mathcal{D}$ contained a lot of information about X.

Information theory primer · Mutual information

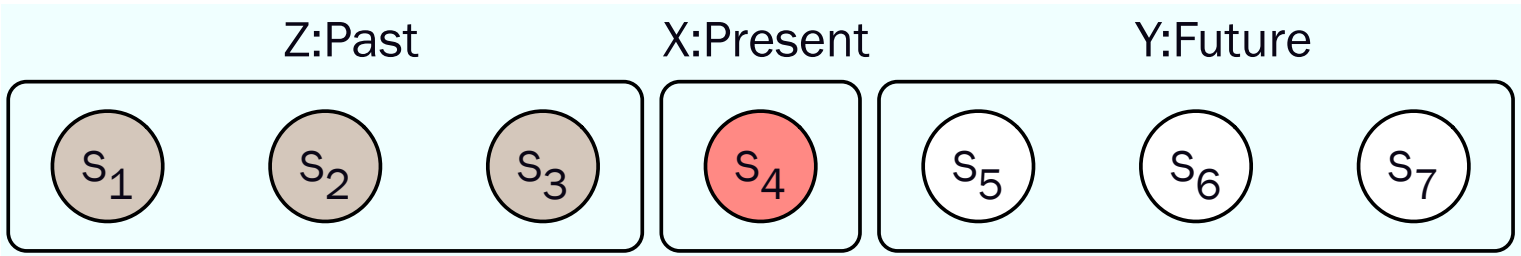
The mutual information between X and Y is the expected amount of information about X in an observation of Y. Can be written in several ways:

$$\begin{aligned} I(X,Y) &= \iint_{\mathcal{X}\times\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)\,p(y)}\mathrm{d}x\mathrm{d}y = D(p_{XY}||p_X\otimes p_Y) \\ &= H(X)+H(Y)-H(X,Y) \\ &= H(X)-H(X|Y) \end{aligned}$$

Interpretations: (1) divergence between joint and product of marginals (hence measure of statistical dependency), (2) difference between entropy and conditional entropy (hence reduction of uncertainty).

Information theory in sequences

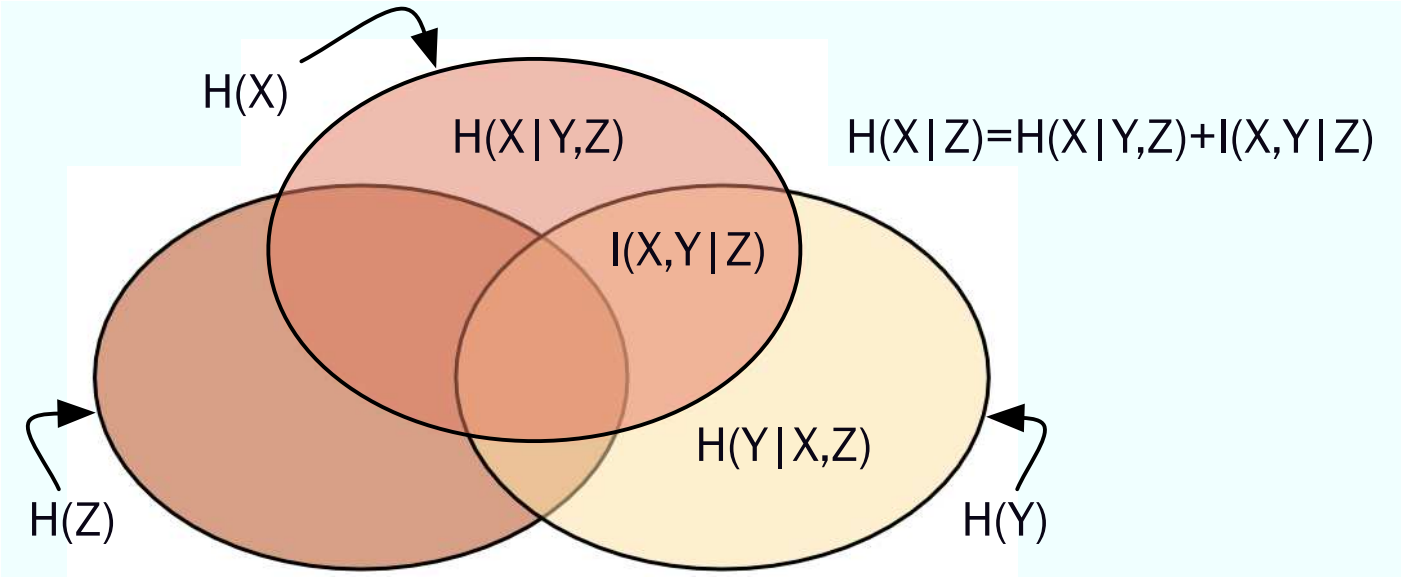
Consider observer receiving elements of a random sequence $S_1, S_2, \ldots$, so that at any time there is an observed past, a 'now', and an unobserved future.



Model is summarised by the observer's probability distribution $p_{XY|Z}$ over the present and future given the past, possibly including involving parameters θ. Consider how the observer's belief state evolves when it learns that $X = x$.

Three-way information measures

Lump together random variables into three sets: $Z = \text{Past}$, $X = \text{Present}$, and $Y = \text{Future}$. Entropies and other information measures all related:



‘Surprise’ based quantities

To obtain first set of 4 measures, we marginalise out the future Y to get distribution for the immediate prediction, $p_{X|Z}$.

1. **Surprisingness**: negative log-probability $\mathcal{L}(x|z) \triangleq -\log p_{X|Z}(x|z)$.

‘Surprise’ based quantities

To obtain first set of 4 measures, we marginalise out the future Y to get distribution for the immediate prediction, $p_{X|Z}$.

1. **Surprisingness**: negative log-probability $\mathcal{L}(x|z) \triangleq -\log p_{X|Z}(x|z)$.
2. Expected surprisingness given current z is the entropy of the predictive distribution, $H(X|Z=z)$: uncertainty about X before the observation is made.

‘Surprise’ based quantities

To obtain first set of 4 measures, we marginalise out the future Y to get distribution for the immediate prediction, $p_{X|Z}$.

1. **Surprisingness**: negative log-probability $\mathcal{L}(x|z) \triangleq -\log p_{X|Z}(x|z)$.
2. Expected surprisingness given current z is the entropy of the predictive distribution, $H(X|Z=z)$: uncertainty about X before the observation is made.
3. Expectation over the possible pasts, i.e. over $Z|X=x$: the average in-context surprisingness of the symbol x, a sort static analysis of the model which picks out which are the most significant states in the state space.

‘Surprise’ based quantities

To obtain first set of 4 measures, we marginalise out the future Y to get distribution for the immediate prediction, $p_{X|Z}$.

1. **Surprisingness**: negative log-probability $\mathcal{L}(x|z) \triangleq -\log p_{X|Z}(x|z)$.
2. Expected surprisingness given current z is the entropy of the predictive distribution, $H(X|Z=z)$: uncertainty about X before the observation is made.
3. Expectation over the possible pasts, i.e. over $Z|X=x$: the average in-context surprisingness of the symbol x, a sort static analysis of the model which picks out which are the most significant states in the state space.
4. Expectation over both X and Z is the entropy rate $H(X|Z)$ according to the observer’s current model.

Predictive information

Second set of 4 measures based on amount of information an observation $X=x$ carries *about* about the unobserved future Y, *given* that we already know the past $Z=z$. Define this as the KL divergence between prior and posterior distributions over future:

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

where

$$p_{Y|Z=z}(y) = \int p_{XY|Z=z}(x,y)\,\mathrm{d}x.$$

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.

Predictive information based quantities

1. *Instantaneous predictive information rate* (IPIR) is just $\mathcal{I}(x|z)$.

Predictive information based quantities

1. *Instantaneous predictive information rate* (IPIR) is just $\mathcal{I}(x|z)$.
2. Expectation $\mathrm{E}_{X|Z=z}\mathcal{I}(X|z)$ is the amount of new information about the future we expect to receive from the next observation. Useful for directing attention towards the next event even before it happens?

Predictive information based quantities

1. *Instantaneous predictive information rate* (IPIR) is just $\mathcal{I}(x|z)$.
2. Expectation $\mathrm{E}_{X|Z=z}\mathcal{I}(X|z)$ is the amount of new information about the future we expect to receive from the next observation. Useful for directing attention towards the next event even before it happens?
3. Expectation over possible pasts, $\mathrm{E}_{Z|X=x}\mathcal{I}(x|Z)$, gives the average 'informativeness' (significance?) of each value in the state space of X. Informative states might tend to appear as 'onset' states, or as the 'foreground' against a 'background' of less informative states.

Predictive information based quantities

1. *Instantaneous predictive information rate* (IPIR) is just $\mathcal{I}(x|z)$.
2. Expectation $\mathrm{E}_{X|Z=z}\mathcal{I}(X|z)$ is the amount of new information about the future we expect to receive from the next observation. Useful for directing attention towards the next event even before it happens?
3. Expectation over possible pasts, $\mathrm{E}_{Z|X=x}\mathcal{I}(x|Z)$, gives the average 'informativeness' (significance?) of each value in the state space of X. Informative states might tend to appear as 'onset' states, or as the 'foreground' against a 'background' of less informative states.
4. Expectation over both X and Z is the *average predictive information rate* (APIR), the average rate at which new information arrives about the future. Reduces to $I(X, Y|Z) = H(Y|Z) - H(Y|X, Z)$.

Information about model parameters

A final measure can be obtained by considering an observer with a paremetric model where parameters are learned on-line.

Observer's belief state includes a probability distribution over the parameters Θ, e.g. $p_{\Theta|Z=z}(\theta)$ is the probability assigned to θ given observed past $Z=z$.

Information about model parameters

A final measure can be obtained by considering an observer with a paremetric model where parameters are learned on-line.

Observer's belief state includes a probability distribution over the parameters Θ, e.g. $p_{\Theta|Z=z}(\theta)$ is the probability assigned to θ given observed past $Z=z$.

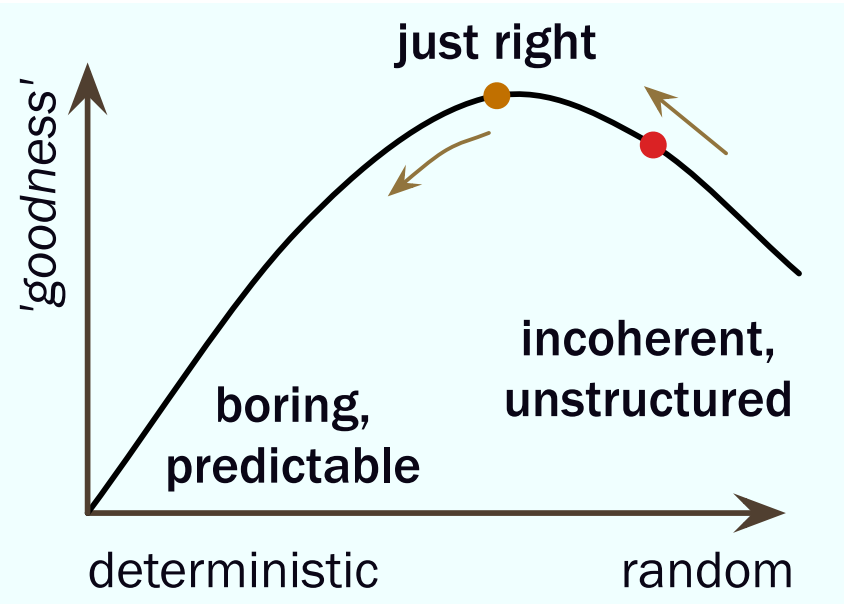
Each observation causes a revision of that belief state and hence supply information about the parameters, quantified as the KL divergence between prior and posterior distributions

$$D(p_{\Theta|X=x,Z=z}||p_{\Theta|Z=z}).$$

We call this the 'model information rate'.

Complexity and aesthetics

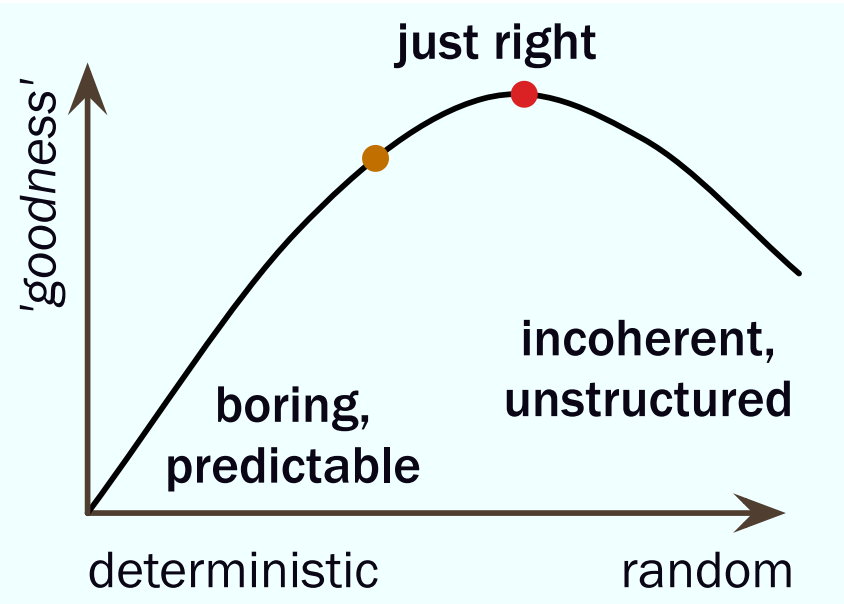
Studies looking into the relationship between stochastic complexity (usually measured as entropy or entropy rate) and aesthetic value, reveal an inverted 'U' shaped curve [Ber71]. (Also, Wundt curve [Wun97]). Repeated exposure tends to move stimuli leftwards.



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.

Complexity and aesthetics

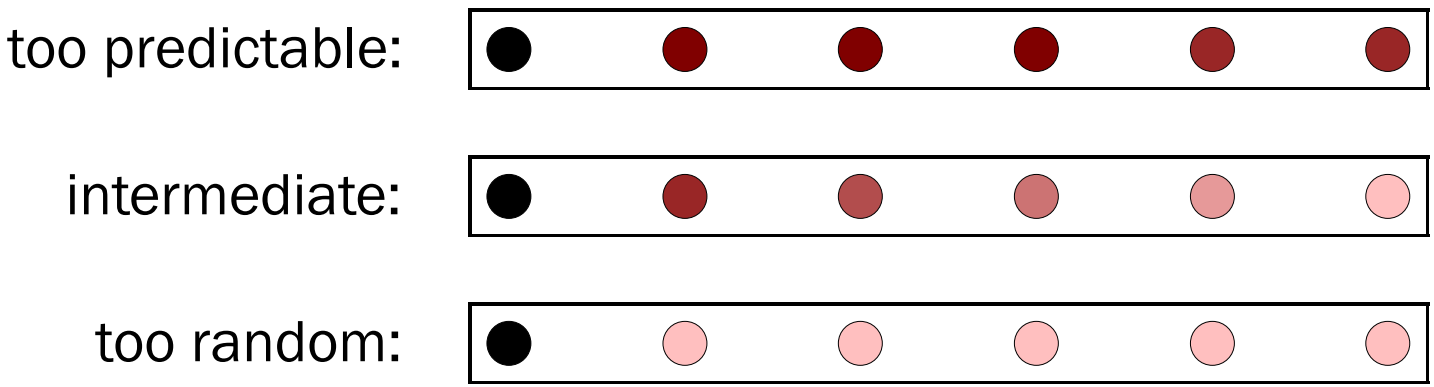
Studies looking into the relationship between stochastic complexity (usually measured as entropy or entropy rate) and aesthetic value, reveal an inverted 'U' shaped curve [Ber71]. (Also, Wundt curve [Wun97]). Repeated exposure tends to move stimuli leftwards.



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.

APIR as a measure of interestingness

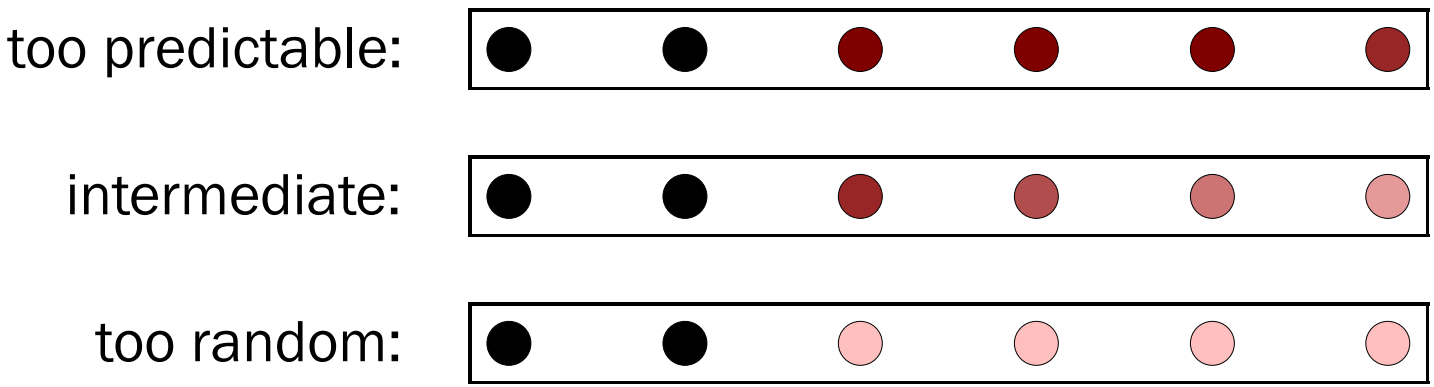
The predictive information rate incorporates a similar balance automatically: is maximal for sequences which are neither deterministic nor totally uncorrelated across time.



(Black: *observed*; red: *unobserved*; paler: *greater uncertainty*.)
Our interpretation: Things are 'interesting' or at least 'salient' when each new part supplies new information about parts to come.

APIR as a measure of interestingness

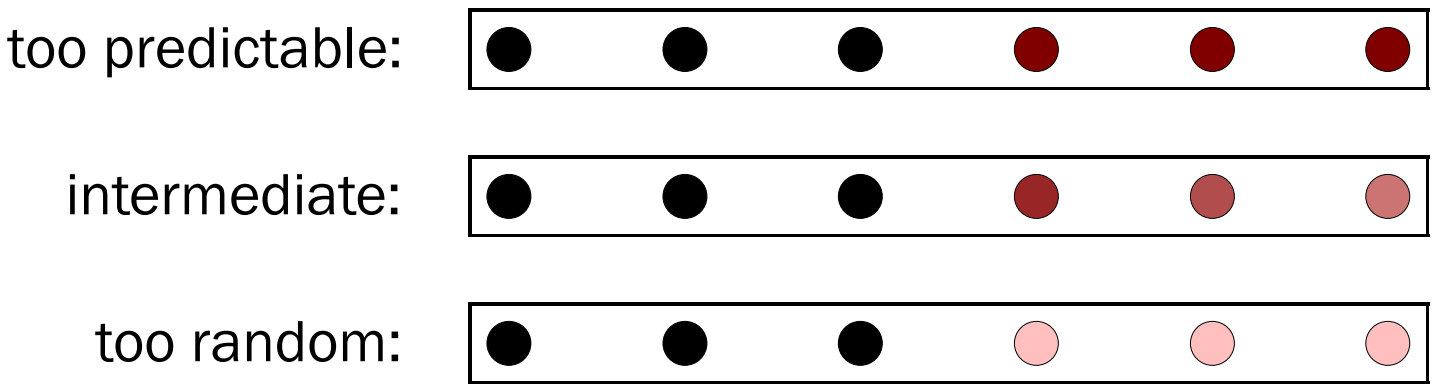
The predictive information rate incorporates a similar balance automatically: is maximal for sequences which are neither deterministic nor totally uncorrelated across time.



(Black: *observed*; red: *unobserved*; paler: *greater uncertainty*.)
Our interpretation: Things are ‘interesting’ or at least ‘salient’ when each new part supplies new information about parts to come.

APIR as a measure of interestingness

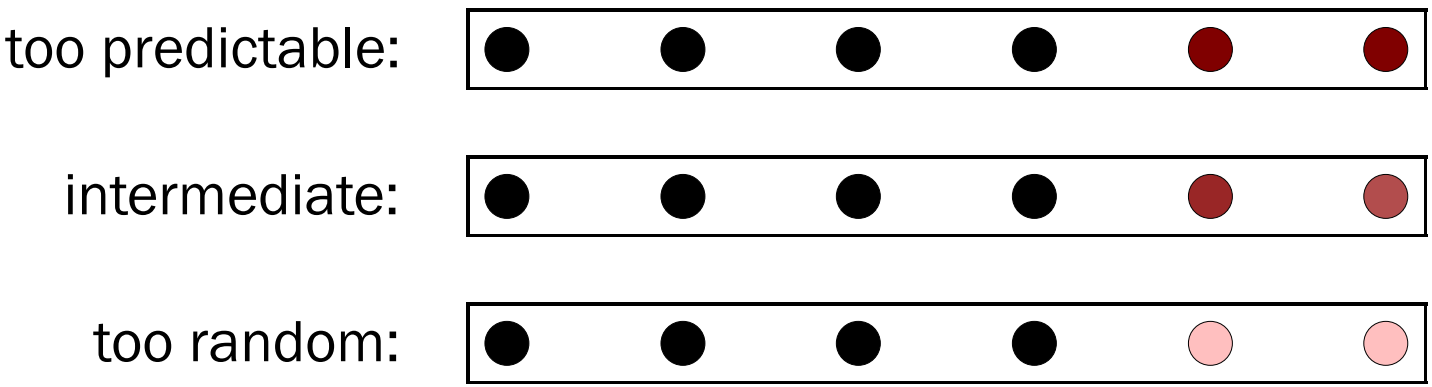
The predictive information rate incorporates a similar balance automatically: is maximal for sequences which are neither deterministic nor totally uncorrelated across time.



(Black: *observed*; red: *unobserved*; paler: *greater uncertainty*.)
Our interpretation: Things are ‘interesting’ or at least ‘salient’ when each new part supplies new information about parts to come.

APIR as a measure of interestingness

The predictive information rate incorporates a similar balance automatically: is maximal for sequences which are neither deterministic nor totally uncorrelated across time.



(Black: *observed*; red: *unobserved*; paler: *greater uncertainty*.)
Our interpretation: Things are ‘interesting’ or at least ‘salient’ when each new part supplies new information about parts to come.

Outline

Markov chains · Definitions I

Now we'll look at information dynamics in one of the simplest possible models, a Markov chain.

Let S be a Markov chain with state space $\{1,\ldots,N\}$ such that S_t is the random variable representing the t^{th} element of the sequence.

Model is parameterised by a transition matrix $a \in \mathbb{R}^{N\times N}$, that is $p(S_{t+1}=i|S_t=j) = a_{ij}$. We require stationarity so we set distribution for initial element S_1 to the equilibrium distribution: $p(S_1=i) = \pi^a_i$ where π^a is a column vector satisfying $a\pi^a = \pi^a$. (To ensure uniqueness of equilibrim distribution, Markov chain must also be irreducible.)

Markov chains · Definitions II

Markov dependency structure means that, for computing dynamic information measures, 'past' and 'future' at time t can be collapsed down to the previous and next elements; i.e., we can set $Z = S_{t-1}$, $X = S_t$, and $Y = S_{t+1}$.

Information measures are expressed (next slide) in terms of 'time-reversed' transition matrix:

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

and the entropy rate

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

Information measures

Surprise-based measures 1–4:

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

(Over- and under-bars denote expectation over S_t and S_{t-1}.)

Information measures

Surprise-based measures 1–4:

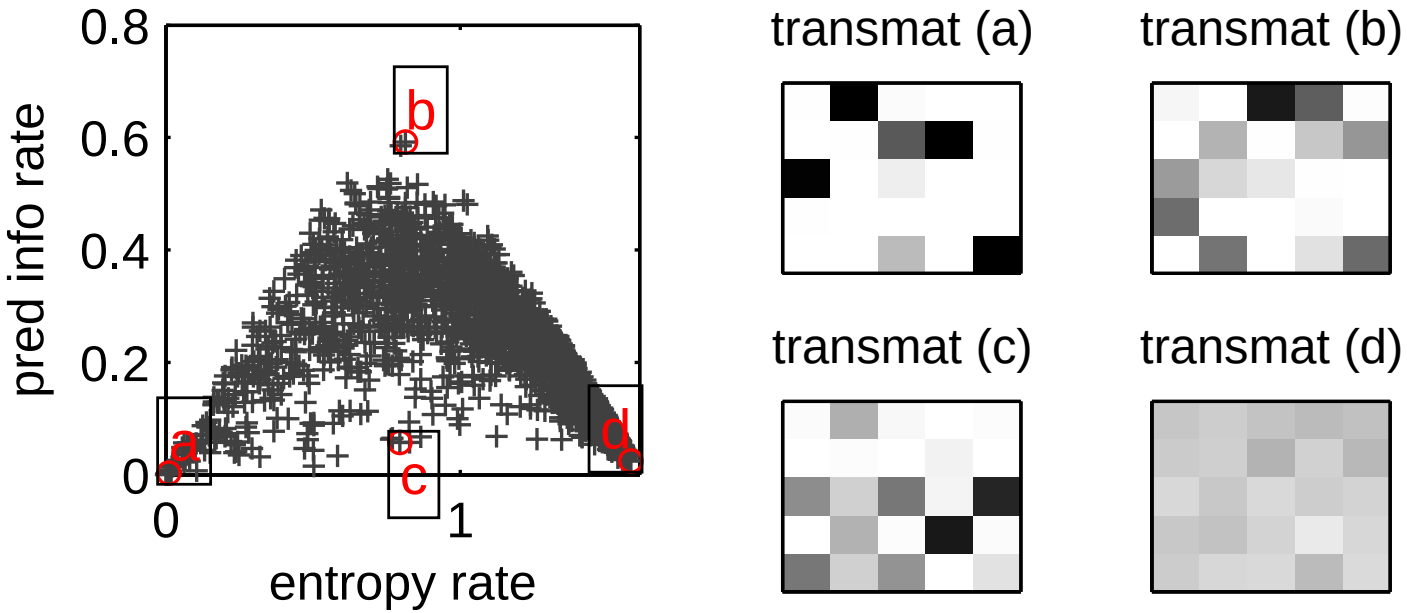
$$
\begin{aligned}
\mathcal{L}(i|j) &= -\log p(S_t{=}j|S_{t-1}{=}i) &&= -\log a_{ij} \\
\overline{\mathcal{L}}(j) &= \mathrm{E}_{\,i\sim S_t|S_{t-1}=j}\mathcal{L}(i|j) &&= \textstyle\sum_{i=1}^{N} a_{ij}\mathcal{L}(i|j) \\
\underline{\mathcal{L}}(i) &= \mathrm{E}_{\,j\sim S_{t-1}|S_t=i}\mathcal{L}(i|j)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

Predictive information-based measures 1–4:

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

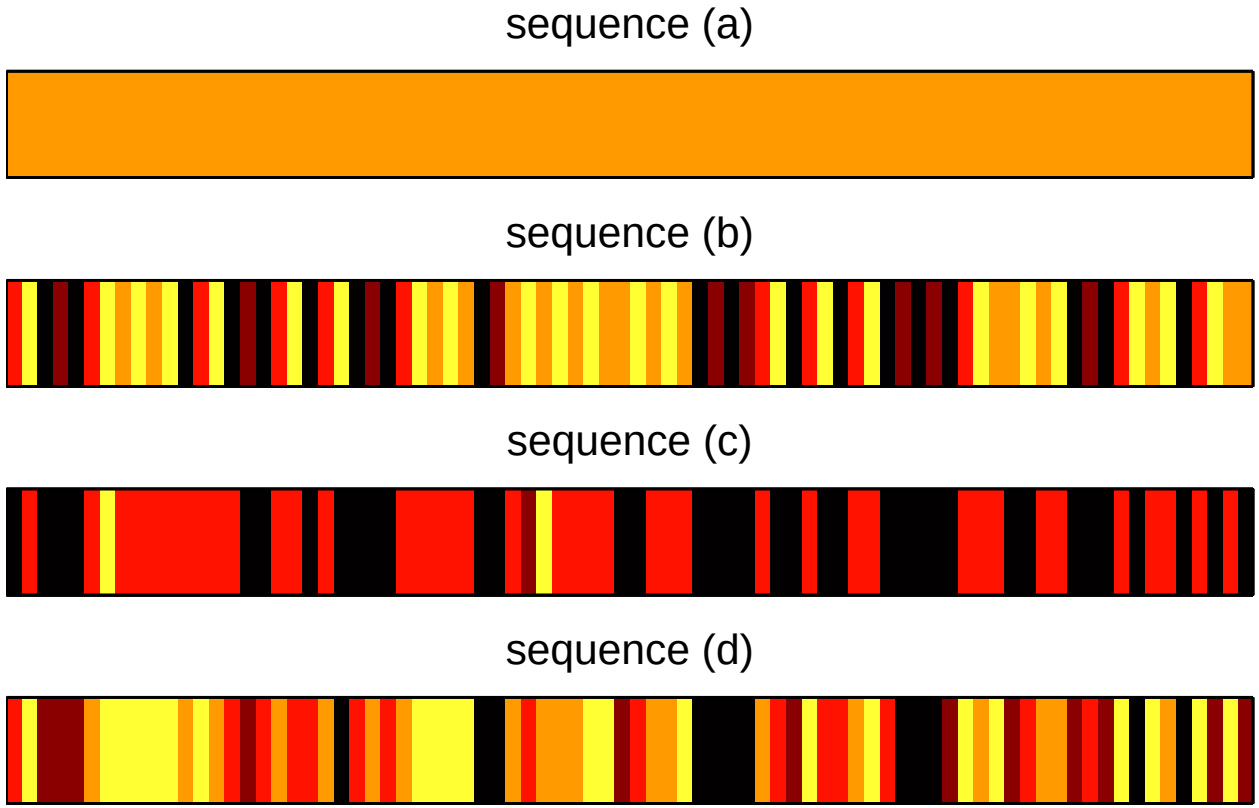
(Over- and under-bars denote expectation over S_t and S_{t-1}.)

Entropy rate and APIR in Markov chains



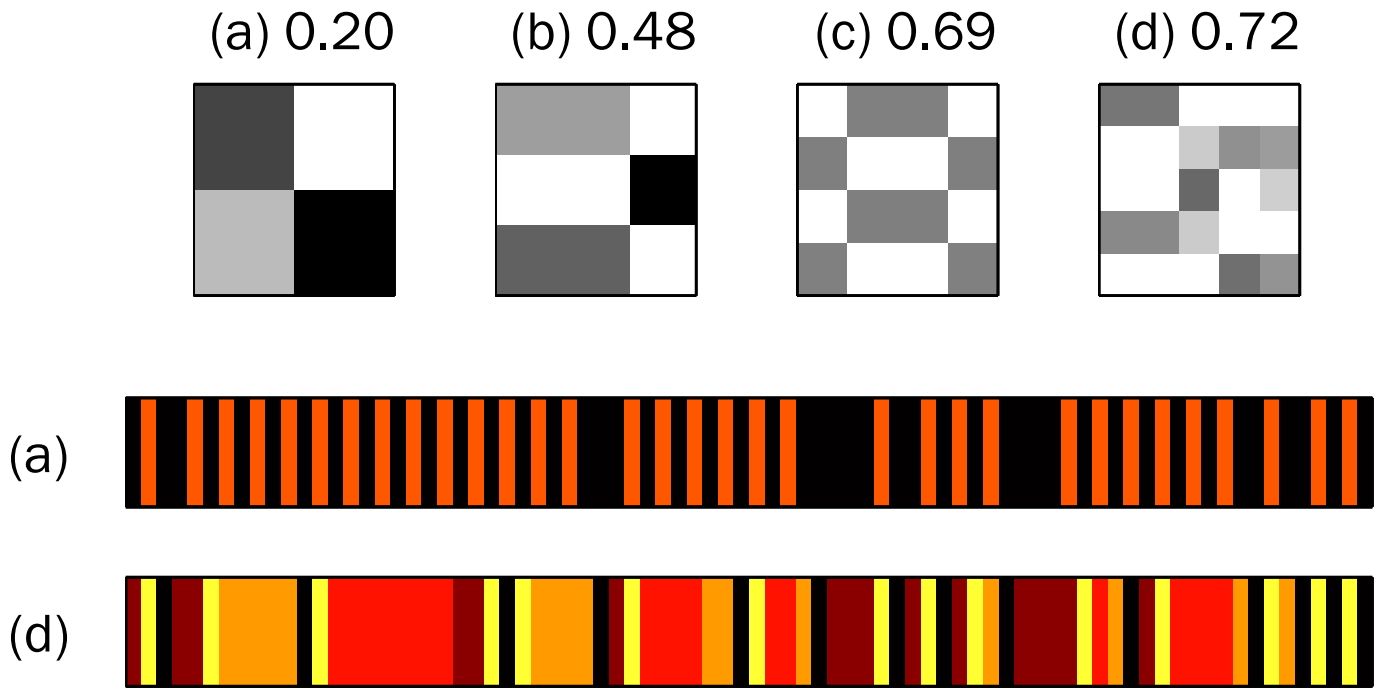
For given N, entropy rate varies between 0 (deterministic sequence) and $\log N$ when $a_{ij} = 1/N$ for all i, j. Space of transition matrices explored by generating them at random and plotting entropy rate vs APIR. (Note inverted 'U' relationship).

Sequences with different APIR



Sequence (a) is repetition of state 4 (see transmat (a) on previous slide). System (b) has the highest APIR.

Direct optimisation of APIR



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.

Outline

Bialek *et al*'s 'Predictive information'

Bialek *et al* [BNT01] consider the entropy of a segment of random process of duration T, which, (given stationarity) will be a function of T alone, say $\mathcal{S}(T)$. This will increase with T, tending towards a linear growth at rate equal to entropy rate of process.

Bialek *et al*'s 'Predictive information'

Bialek *et al* [BNT01] consider the entropy of a segment of random process of duration T, which, (given stationarity) will be a function of T alone, say $\mathcal{S}(T)$. This will increase with T, tending towards a linear growth at rate equal to entropy rate of process.

Mutual information between two adjacent segments, of duration T and T' can be expressed in terms of $\mathcal{S}$. Bialek *et al* define their predictive information as the limit of this as T' tends to infinity:

$$I_{\text{pred}}(T) = \lim_{T' \to \infty} \mathcal{S}(T) + \mathcal{S}(T') - \mathcal{S}(T + T'). \tag{1}$$

Bialek *et al*'s 'Predictive information'

Bialek *et al* [BNT01] consider the entropy of a segment of random process of duration T, which, (given stationarity) will be a function of T alone, say $\mathcal{S}(T)$. This will increase with T, tending towards a linear growth at rate equal to entropy rate of process.

Mutual information between two adjacent segments, of duration T and T' can be expressed in terms of $\mathcal{S}$. Bialek *et al* define their predictive information as the limit of this as T' tends to infinity:

$$I_{\text{pred}}(T) = \lim_{T' \to \infty} \mathcal{S}(T) + \mathcal{S}(T') - \mathcal{S}(T + T'). \tag{1}$$

Behaviour as $T \to \infty$ (finite limit, logarithmic or power-law growth) characterises stochastic complexity of process.

Bialek *et al*'s 'Predictive information'

Bialek *et al* [BNT01] consider the entropy of a segment of random process of duration T, which, (given stationarity) will be a function of T alone, say $\mathcal{S}(T)$. This will increase with T, tending towards a linear growth at rate equal to entropy rate of process.

Mutual information between two adjacent segments, of duration T and T' can be expressed in terms of $\mathcal{S}$. Bialek *et al* define their predictive information as the limit of this as T' tends to infinity:

$$I_{\text{pred}}(T) = \lim_{T' \to \infty} \mathcal{S}(T) + \mathcal{S}(T') - \mathcal{S}(T + T'). \tag{1}$$

Behaviour as $T \to \infty$ (finite limit, logarithmic or power-law growth) characterises stochastic complexity of process.

$I_{\text{pred}}(T)$ is a *global* measure which applies to process as a whole, not to specific instants within a realisation: hence it doesn't give a dynamic analysis of observed sequences.

Dubnov's 'information rate'

Dubnov [Dub06] proposes an 'information rate' (IR) which, in our notation, is $I(S_t, S_{-\infty:t-1})$, i.e. the mutual information between the past and the present.

For a Markov chain, this reduces to $\mathcal{H}_0(\pi^a) - \dot{\mathcal{H}}(a)$, where $\mathcal{H}_0(\pi^a)$ is the entropy of the equilibrium distribution π^a.

Dubnov's 'information rate'

Dubnov [Dub06] proposes an 'information rate' (IR) which, in our notation, is $I(S_t, S_{-\infty:t-1})$, i.e. the mutual information between the past and the present.

For a Markov chain, this reduces to $\mathcal{H}_0(\pi^a) - \dot{\mathcal{H}}(a)$, where $\mathcal{H}_0(\pi^a)$ is the entropy of the equilibrium distribution π^a.

Dubnov argues that this has the 'inverted-U' characteristic, but for Markov chains at least, the effect is not what we expect: Dubnov's IR *is* zero for sequences of independent events, but maximal IR is reached by simultaneously minimising the entropy rate and maximising the entropy of π^a. Corresponds to uniform π^a but deterministic transitions thereafter.

Deterministic cycling through states will have this property. APIR is zero in these cases.

Other related work

Information gained about model parameters (measured as the KL divergence between prior and posterior distributions) is equivalent to **Itti and Baldi's 'Bayesian surprise'** [IB05].

Other related work

Information gained about model parameters (measured as the KL divergence between prior and posterior distributions) is equivalent to **Itti and Baldi's 'Bayesian surprise'** [IB05].

Eerola *et al* [ETK02] emphasise the need for dynamic probability models when judging uncertainty and predictability in music. They also describe experimental methods for assessing these quantities in human listeners. They do not explore multiple information measures or consider predictive information.

Other related work

Information gained about model parameters (measured as the KL divergence between prior and posterior distributions) is equivalent to **Itti and Baldi's 'Bayesian surprise'** [IB05].

Eerola *et al* [ETK02] emphasise the need for dynamic probability models when judging uncertainty and predictability in music. They also describe experimental methods for assessing these quantities in human listeners. They do not explore multiple information measures or consider predictive information.

Levy and Jaeger [LJ07] study '**information density' in spoken langugage** using surprisingness and show speakers often choose their words in order to acheive a constant information rate.

Outline

Material and Methods

We took two pieces of minimalist music by Philip Glass, *Two Pages* (1969) and *Gradus* (1968). Both monophonic and isochronous, so representable very simply as a sequence of symbols (notes), one symbol per beat, yet remain ecologically valid examples of ‘real’ music.

We use an elaboration of the Markov chain model—not necessarily a good model *per se*, but that wasn’t the point of the experiment. Markov chain model enables exact analysis without approximations.

Time-varying transition matrix model

We allow transition matrix to vary slowly with time to track changes in the sequence structure. Hence, observer's belief state includes a probabilitiy distribution over transition matrices; we choose a product of Dirichlet distributions:

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

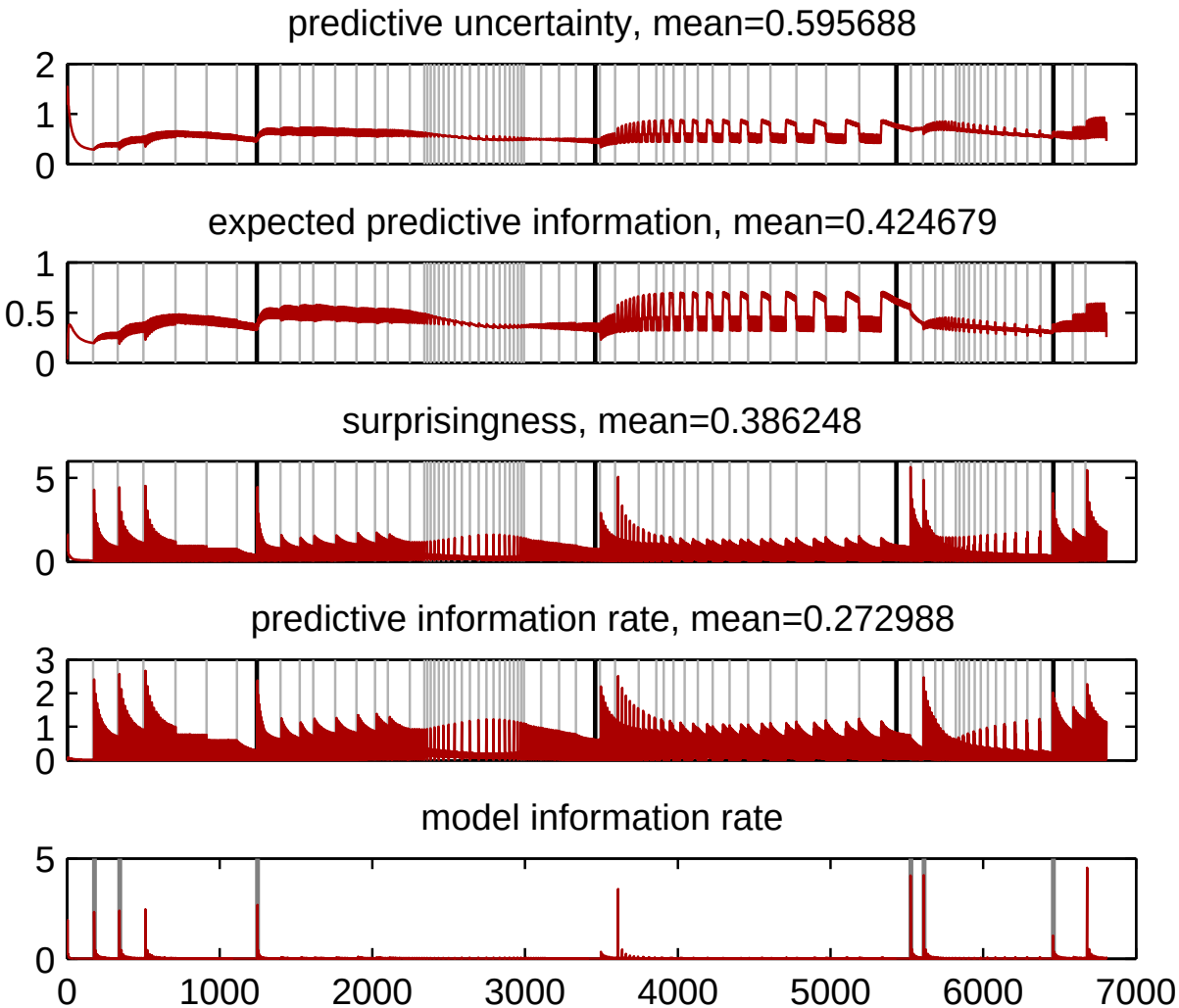
where $a_{:j}$ is j^{th} column of a and θ is an $N \times N$ parameter matrix.

At each time step, distribution first *spreads* under mapping

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

to model possibility that transition matrix has changed ($\beta = 2500$ in our experiments). Then it *contracts* due to new observation providing fresh evidence about transition matrix.

Two Pages · Results



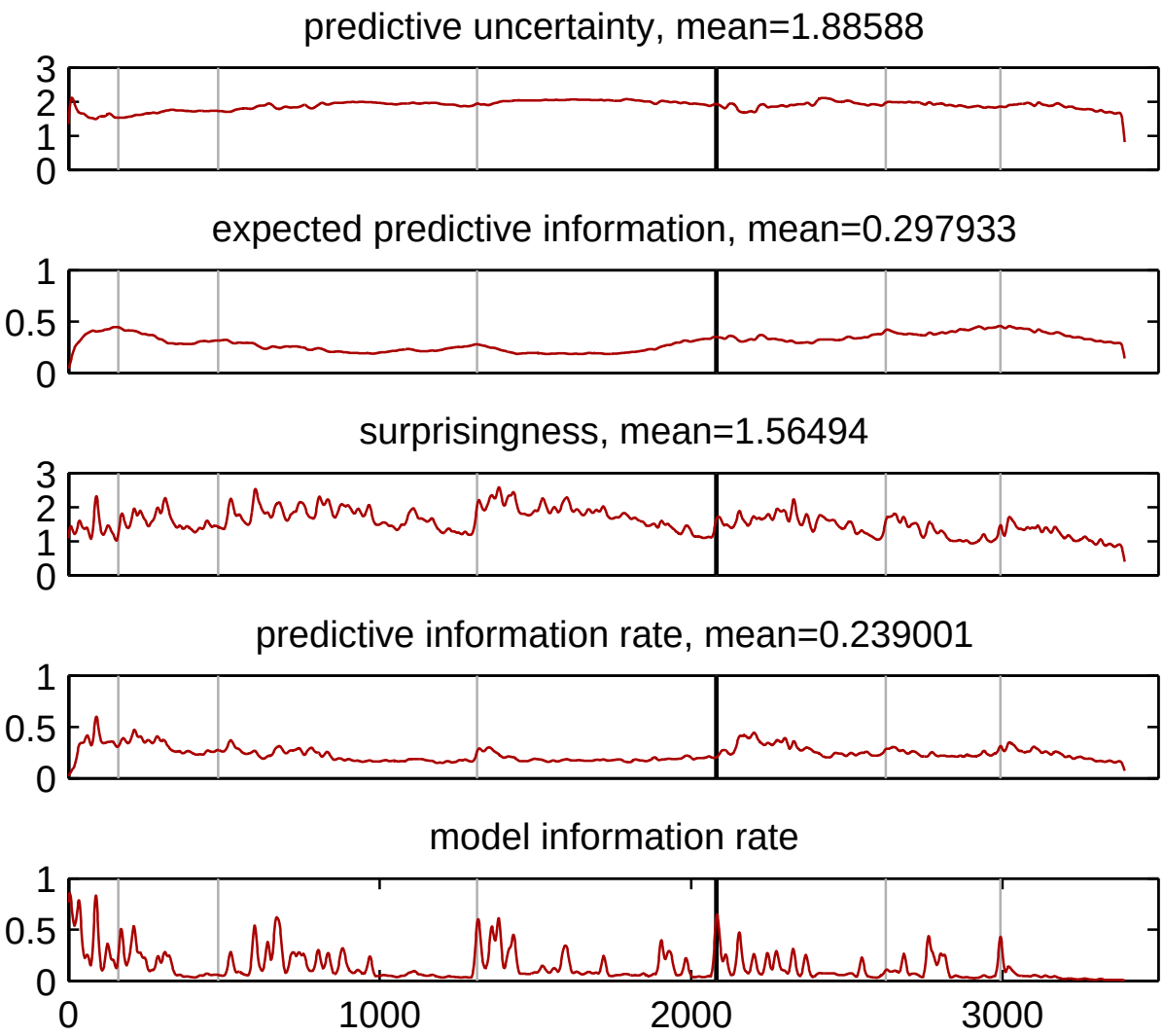
Thick lines: part boundaries as indicated by Glass; **grey lines (top four panels):** changes in the melodic 'figures'; **grey lines (bottom panel):** six most surprising moments chosen by expert listenter.

Two Pages · Discussion

Correspondence between the information measures and the structure of the piece is quite close. Good agreement between the six ‘most surprising moments’ chosen by expert listener and model information signal.

What appears to be an error in the detection of the major part boundary (between events 5000 and 6000) 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 · Results



Thick lines: part boundaries as indicated by the composer. **Grey lines:** segmentation by expert listener. Note: traces smoothed with Gaussian window about 16 events wide.

Gradus · Discussion

Gradus is much less systematically structured than *Two Pages*, and relies more on the conventions of tonal music, which are not represented the model.

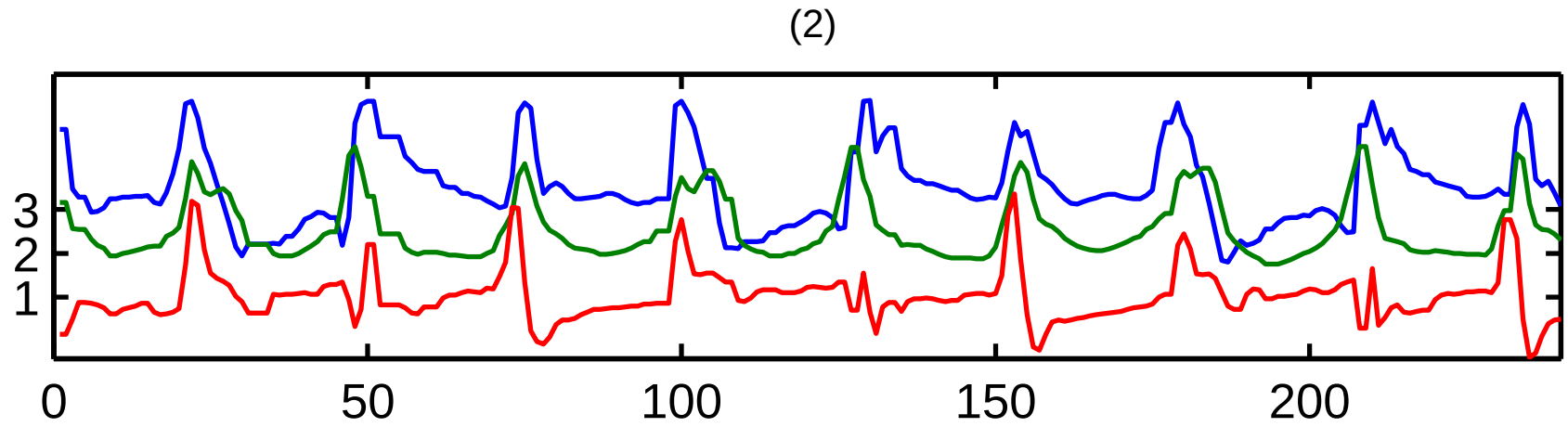
For example initial transition matrix is uniform, which does not correctly represent prior knowledge about tonal music.

Information dynamic analysis does not give such a clear picture of the structure; but some of the fine structure can be related to specific events in the music (see Pearce and Wiggins 2006).

Outline

Application to gesture recognition

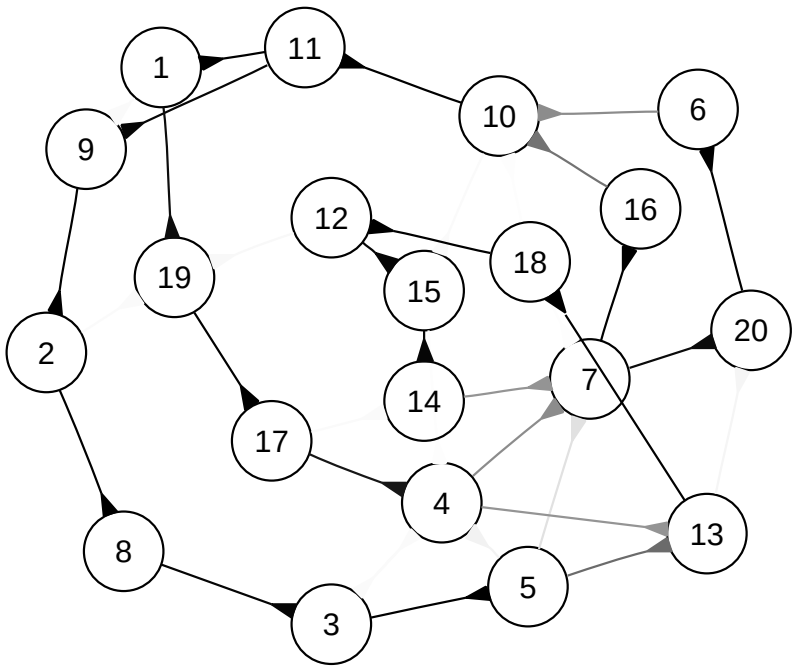
Problem is to detect and classify the gestures made by a conductor has he or she beats time, i.e., the events which mark the beat times.



Data consists of 3 accelerometer signals from a Nintendo Wii controller. Can we detect and localise the relevant events by looking for moments of high predictive information rate?

HMM fitted to Wii data

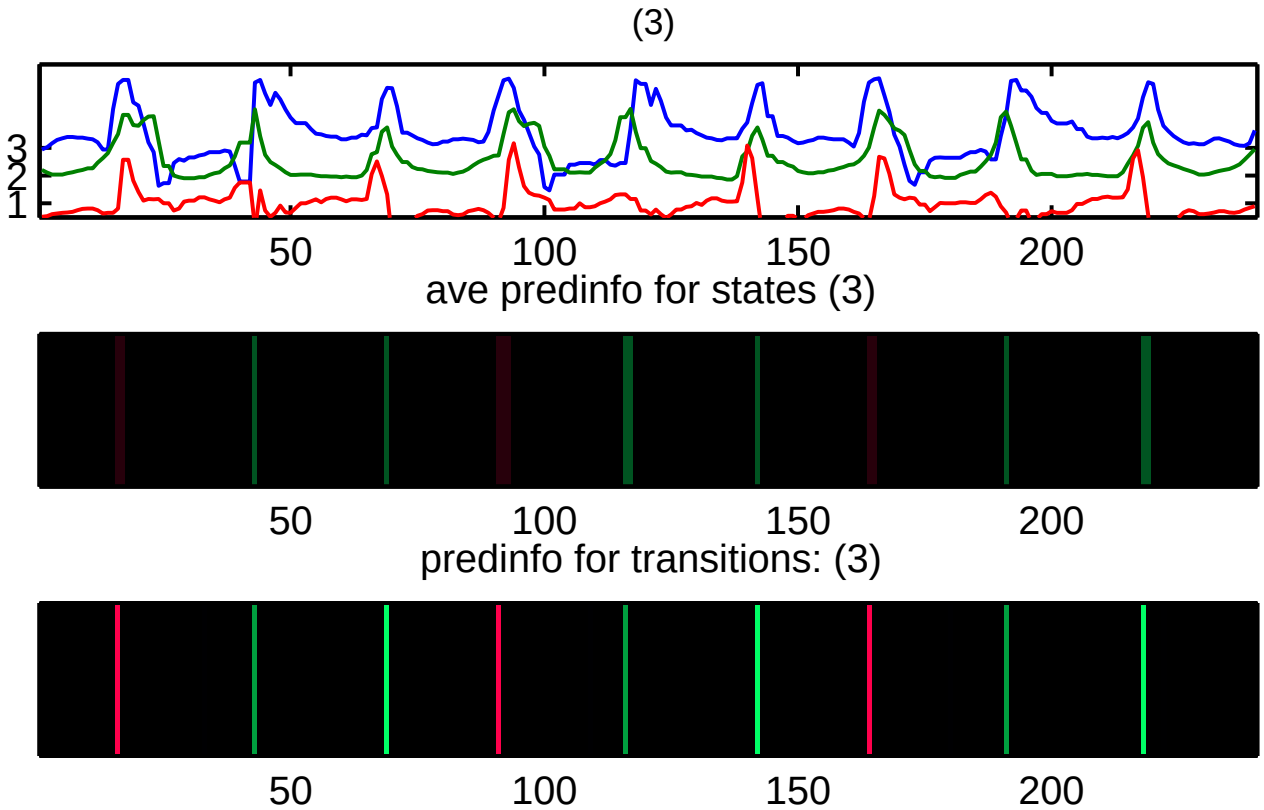
One option is to use an HMM to encode the real-valued data as a sequence of discrete symbols and then use the Markov chain analysis to compute the information rates.



Only a rough approximation of the information dynamics of the HMM as a whole as we are ignoring uncertainty about the hidden state sequence. This graph illustrates a fairly sparse transition matrix resulting from EM learning (self transitions not shown).

Predictive information in HMM state sequence

Middle plot uses brightness to show average informativeness of each state. Bottom plot uses brightness to show predictive information in each transition. (Hue encodes state number.)



Approximations for dealing with latent variables

It was easy to compute information dynamic quantities for fully observed Markov chains, but these are limited in the range of phenomena they can model.

What about more powerful models with latent variables like the HMM we just looked at?

Approximations for dealing with latent variables

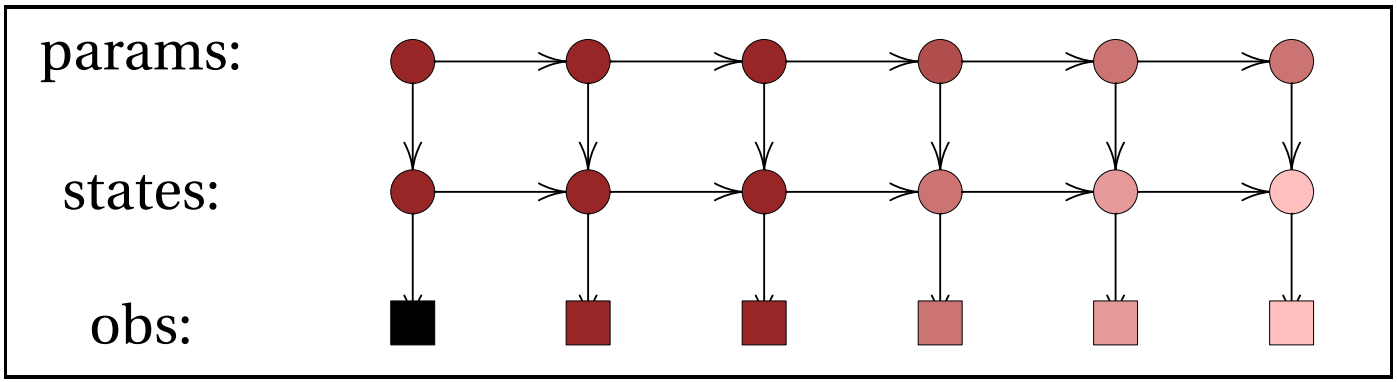
It was easy to compute information dynamic quantities for fully observed Markov chains, but these are limited in the range of phenomena they can model.

What about more powerful models with latent variables like the HMM we just looked at?

One option is to use variational Bayesian methods: use a tractable family of distributions to model observer's beliefs about latent variables. Then we can track entropy and information wrt latent variables almost for free.

Approximations for dealing with latent variables II

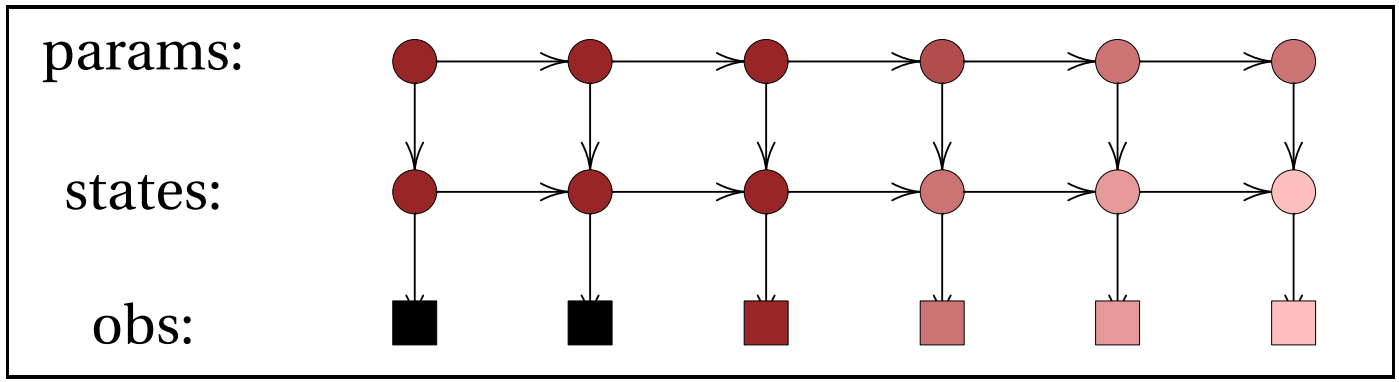
Variational Bayesian filtering algorithms available to deal specifically with online or sequential processing (e.g. [vQ06]).



Above model allows to consider sequential information gain about variables at different *levels* as well as at *times*.

Approximations for dealing with latent variables II

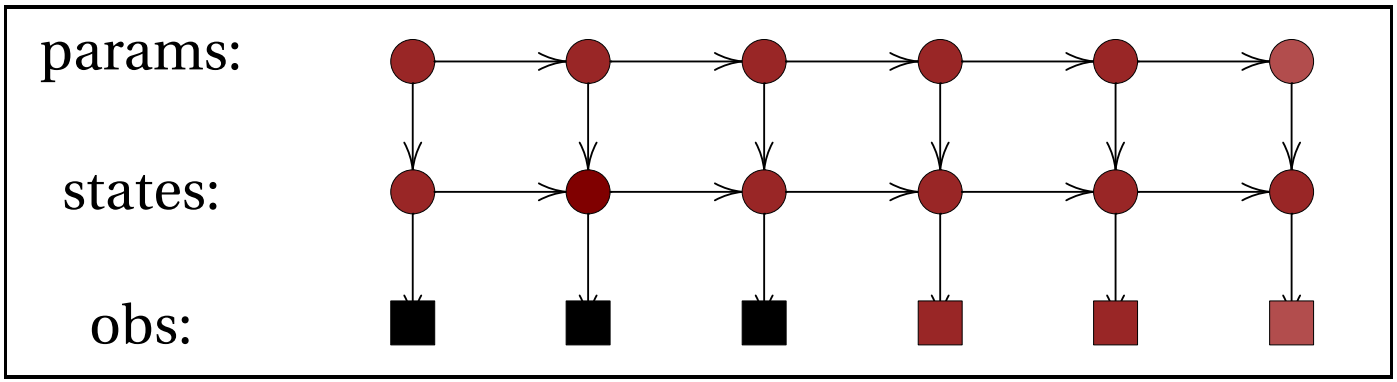
Variational Bayesian filtering algorithms available to deal specifically with online or sequential processing (e.g. [vQ06]).



Above model allows to consider sequential information gain about variables at different *levels* as well as at *times*.

Approximations for dealing with latent variables II

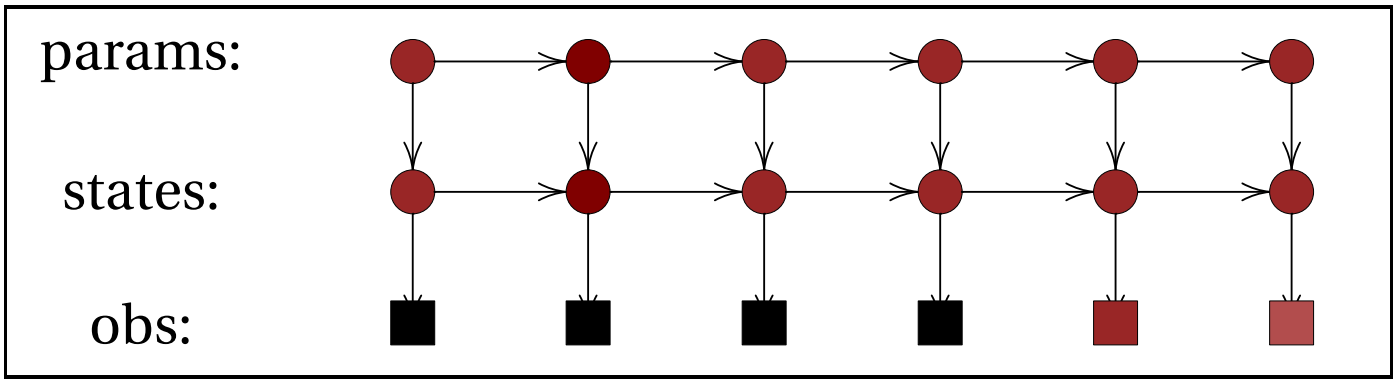
Variational Bayesian filtering algorithms available to deal specifically with online or sequential processing (e.g. [vQ06]).



Above model allows to consider sequential information gain about variables at different *levels* as well as at *times*.

Approximations for dealing with latent variables II

Variational Bayesian filtering algorithms available to deal specifically with online or sequential processing (e.g. [vQ06]).



Above model allows to consider sequential information gain about variables at different *levels* as well as at *times*.

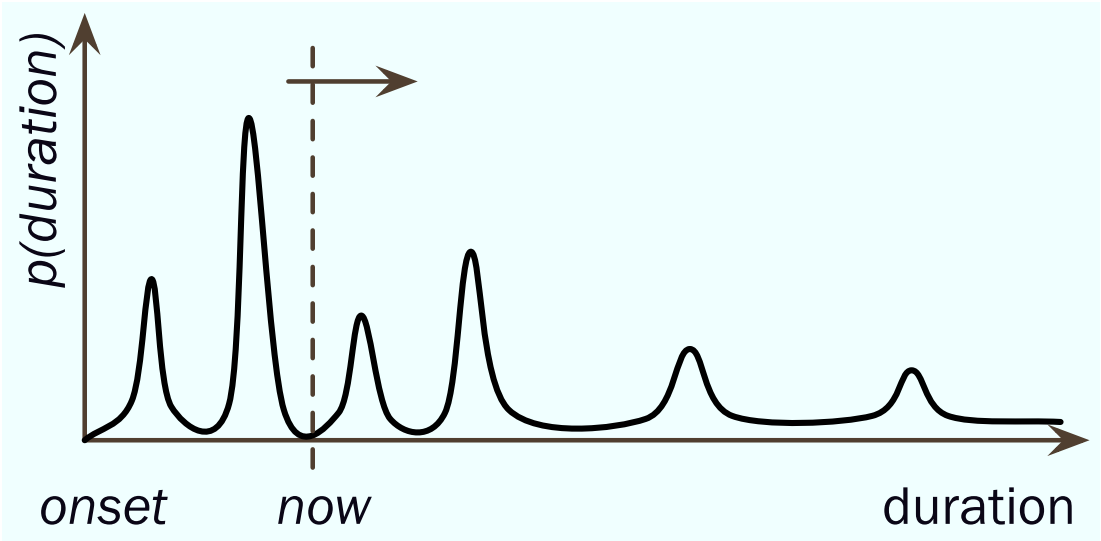
Outline

Summary

- Dynamic, observer-centric information theory.
- Applicable to any dynamic probabilistic model.
- APIR displays inverted-‘U’ with entropy rate.
- Simple analysis for Markov chains
- Still tractable for HMMS using online variational Bayes
- Plausible results when applied to music, but needs more validation.

Future work I

- Consider variable-duration events: if observer expects durations to follow a certain distribution, what is the rate of information arriving while observer *waits* for next event?



Future work II

- Investigation of info-dynamics in HMMs using online variational methods.
- Investigate interaction between learning of intra- and extra-opus style.
- Experiments with human subjects:
  - Relationship between predictive information and ‘interestingness’ or aesthetic value? (The author certainly finds the high APIR processes least maddening to listen to!)
  - Neural correlates of the information measures? Eg, already known that some ERP (eg ERAN [Jen07]) relate to suprise but what about uncertainty, predictive information, and belief revision?

Acknowledgements

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

Bibliography I

- D. E. Berlyne.
  Aesthetics and Psychobiology.
  Appleton Century Crofts, New York, 1971.
- William Bialek, Ilya Nemenman, and Naftali Tishby.
  Predictability, complexity, and learning.
  Neural Computation, 13:2409—2463, 2001.
- J. E. Cohen.
  Information theory and music.
  Behavioral Science, 7(2):137–163, 1962.
- Richard T. Cox.
  Probability, frequency and reasonable expectation.
  American Journal of Physics, 14:1–13, 1946.

Bibliography II

- Daryl Conklin and Ian H. Witten.

  Multiple viewpoint systems for music prediction.

  Journal of New Music Research, 24(1):51–73, 1995.

- Stephen Davies.

  Philosophical perspectives on music's expressiveness.

  In Patrick N. Juslin and John A. Sloboda, editors, *Music and Emotion — Theory and Research*, chapter 2, pages 23–44. Oxford University Press, 2004.

- Shlomo Dubnov.

  Spectral anticipations.

  Computer Music Journal, 30(2):63–83, 2006.

Bibliography III

- T. Eerola, P. Toiviainen, and C. L. Krumhansl.

  Real-time prediction of melodies: Continous 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)*, Syndney, Australia, 2002. Causal Productions.

- E. Hanslick.

  On the musically beautiful: A contribution towards the revision of the aesthetics of music.

  Hackett, Indianapolis, IN, 1854/1986.

- Laurent Itti and Pierre Baldi.

  Bayesian surprise attracts human attention.

  In *Advances Neural Information Processing Systems (NIPS 2005)*, 2005.

Bibliography IV

- 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.

- S. Jentschke.

  Psychoacoustic influences on the neural correlates of music syntactic processing, 2007.

  NIPS 2007 Workshop on Brain, Music and Cognition.

- Susanne K. Langer.

  Philosophy in a new key.

  Harvard University Press, Cambridge, MA, 1957.

Bibliography V

- 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.

- Leonard B. Meyer.

  Music, the arts and ideas: Patterns and Predictions in Twentieth-century culture.

  University of Chicago Press, 1967.

- Abraham Moles.

  Information Theory and Esthetic Perception.

  University of Illinois Press, 1966.

Bibliography VI

- Eugene Narmour.

  Beyond Schenkerism.

  University of Chicago Press, 1977.

- 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.

- 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.

- D. Stern.

  The Interpersonal World of the Infant.

  Academic Press, London, 1985.

Bibliography VII

- Václav Šmídl and Anthony Quinn.

  The variational bayes approximation in bayesian filtering.

  In *ICASSP 2006*, pages III – 137–140, 2006.

- W. Wundt.

  Outlines of Psychology.

  Englemann, Lepzig, 1897.