Tutorial on Integrated Information, Causal Density and Conscious Level

Section on Information Theory and Integration Information

Adam Barrett

University of Sussex
Sackler Centre for Consciousness Science

Tutorial, ASSC 15, Kyoto 2011
2. Introduction to information theory
**Probability Theory**

\[ X \] Random variable (r.v.)

\[ x \] Outcome of \( X \)

\[ X|Y = y \] Conditional r.v. for \( X \) given that \( Y=y \)

\[ P\{X = x\} \] Probability that \( X=x \) (discrete variable)

\[ P(x) \] Probability density (continuous variable)
Entropy (uncertainty)

Can characterize surprise by the negative of the log-probability:

\[ H(X) = - \sum_x P\{X = x\} \log P\{X = x\} \]
Entropy (uncertainty)

\[ H(X) = - \sum_x P\{X = x\} \log P\{X = x\} \]

Later on: view as information capacity

Desirable properties:

- Additive: \( H(X,Y) = H(X) + H(Y), \quad X,Y \text{ indep} \)

- Maximum for uniform distribution: unbiased coin toss is maximum entropy.
Conditional Entropy and Information

Conditional entropy of X given that Y=y:

\[ H(X|Y = y) = - \sum_x P\{X = x|Y = y\} \log P\{X = x|Y = y\} \]

Information generated about X by knowing Y=y:

\[ I(X|Y = y) = H(X) - H(X|Y = y) \]

`Reduction in uncertainty`
Simple example

Y is on if and only if X is off.

A priori all possibilities equal

\[ P\{Y = 0\} = \frac{1}{2}, \quad P\{Y = 1\} = \frac{1}{2} \quad \Rightarrow \quad H(Y) = \log 2 \]

\[ P\{Y = 0|X = 0\} = 0, \quad P\{Y = 1|X = 0\} = 1, \quad \Rightarrow \quad H(Y|X = 0) = 0 \]

\[ I(Y|X = 0) = \log 2 \]
Conditional Entropy and Information

Information generated about X by knowing Y=y:

\[ I(X|Y = y) = H(X) - H(X|Y = y) \]

Note: Can be negative!

(If distribution very non-uniform)
Average (Mutual) Information

Conditional entropy is on average less than the unconditional entropy. Thus define

\[ H(X|Y) = \sum_y P(Y = y)H(X|Y = y) \]

Then, overall information Y contains about X is:

\[ I(X; Y) = H(X) - H(X|Y) \]
Mutual Information (exercises)

\[ I(X; Y) = H(X) - H(X|Y) \]

Useful Exercise: Show

1. \( I(X; Y) \geq 0 \), with equality if and only if \( X \) and \( Y \) independent

2. \( I(X; Y) = H(X) + H(Y) - H(X, Y) \)

3. \( I(X; Y) = I(Y; X) \), (symmetry)

4. \( I(X; Y) \leq \min\{H(X), H(Y)\} \)
Continuous variables and time series

Strictly, entropy for continuous variable is infinite.

However, replacing sum with integral defines sensible quantity:

`Differential entropy’

\[ H(X) = - \int P(x) \log P(x) \, dx \]

Not always positive, but informational quantities have same properties as for discrete case
Gaussian case

In absence of advanced techniques, common to attempt Gaussian approximation.

Regularity of distribution means entropy just depends on variance:

$$P(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ \frac{(x - \mu)^2}{\sigma^2} \right\}$$

Ex. Show that $H(X) \propto \log(\sigma)$
**Multivariate Gaussian**

Entropy depends on determinant of covariance matrix, i.e. `volume of cloud`

\[
COV(\hat{X}) = \begin{pmatrix}
Var(X_1) & Cov(X_1, X_2) \\
Cov(X_2, X_1) & Var(X_2)
\end{pmatrix}
\]

\[
H(X) \propto \log(\det(COV(\hat{X})))
\]
Information between Gaussians

Consider 2 Gaussians $X$ and $Y$.

Relationship has to be linear:

$$X = aY + \varepsilon \quad \varepsilon \text{ indep of } X$$

$$I(X; Y) = H(X) - H(X|Y)$$

$$= H(X) - H(\varepsilon) \quad \text{Conditional entropy indep of value } y \text{ of } Y$$

$$\propto \log(\text{Var}(X)) - \log(\text{Var}(\varepsilon))$$

$$\text{Var}(\varepsilon) = \text{Var}(X) - \text{Cov}(X, Y)^2 [\text{Var}(Y)]^{-1}$$

MESSAGE: formula from variance and covariance
3. Exploration of integrated information ($\Phi$)
Conjoined differentiation and integration

Proposed quantitative measures of conscious level:

- **Neural complexity** (Tononi, Sporns, Edelman, 1994)
  Initial measure. Based only on instantaneous correlations.

- **Information integration (Φ)** (Tononi, ‘04, ‘08, Barrett & Seth ‘11)

- **Causal density (cd)** (Seth, 2005, 2008)
  - Φ (2008,2011) and cd consider directed interactions.
  - cd utilizes Granger causality, which could also be more useful than synchrony for studying integration.
**Integrated information theory of consciousness**

Tononi- “Consciousness is integrated information.”

**Integrated information**: the extent to which the whole system generates more information than the sum of its parts

- The theory provides a measure of conscious level, $\Phi$.
- Conscious content (qualia) determined by informational relationships between elements of the system.

What Information?

1. Tononi: Micro-level

- Binary functional units.
- Consider information generated every time there is a state transition.
- Information current state tells you about the past state.
- The more past states that are ruled out by the current state, the greater the information generated.

Technically measured using Shannon information, and probability theory, from a complete generative model of the system.
Information current state tells you about previous state, on average during a short epoch.

Agnostic on strict identity of C with Φ. Instead explore potential for correlation with C in the brain.
Possible interpretation as information integrated about activity that entered the core $t$ ms ago.

Uses time series analysis

[Barrett, Seth 2011]
**Integration**

- **Integration:** whole is greater than sum of parts.

\[
\Phi = \text{Inf(whole)} - \text{Inf(Part1)} - \text{Inf(Part2)}
\]

(parts chosen with weakest link between them)

---


Tononi & Balduzzi (2008), *PLoS Comp Biol*
Effective information

Effective information with respect to a bipartition

\[ \varphi[X; \tau, \{M, N\}] = I(X_{t-\tau}; X_t) - I(M_{t-\tau}; M_t) - I(N_{t-\tau}; N_t) \]

[Barrett, Seth 2011]
Information integration

Effective information with respect to the minimum information bipartition (MIB)

\[ \Phi[X; \tau] = \varphi[X; \tau, B^{MIB}] \]
Normalization

Consider case of binary elements: single element can share only 1 bit of information.

A more equal split (here 5-4) can share 4 bits.

Divide by capacity of partition:

\[
\text{Minimize normalized effective information: } \frac{\phi(M, N)}{\min\{H(M), H(N)\}}
\]
Information integration

Effective information with respect to the minimum information bipartition (MIB)

\[ \Phi[X; \tau] = \varphi[X; \tau, B^{MIB}] \]
Complexes

A **complex** is a system subset that has higher $\Phi$ than any larger system subset that it is contained in.

- Systems can have many overlapping complexes.
- These are where consciousness “resides”.
- Brain (hopefully!) has just 1 high $\Phi$ complex! (Split brain patients may have 2...
Examples

Dynamics given by connectivity matrix:

\[ X_t = A \cdot X_{t-1} + \varepsilon_t \]
Optimal Networks

\[ \Phi = 0.25 \]

\[ \Phi = 0.30 \]
Evidence for information integration

TMS-hdEEG

Ferrarelli, Massimini, Tononi et al. (2005, 2010).
**Auto-regressive reformulation**

For Gaussian data empirical entropy is:

- is easy to compute,
- is based on regression errors.

Why not generalize the algorithm to create a new measure?

\[ X^i_{t-\tau} = \sum_j A_{ij} X^j_t + E^i_t \]  

Regress past on future

Compare variance of X before and after regression:

\[ PP(X_{t-\tau} | X_t) = \log \left( \frac{\text{det}(COV(X))}{\text{det}(COV(E))} \right) \]

[Barrett, Seth 2011]
Auto-regressive $\Phi$

$$AR\varphi[X; \tau, \{M, N\}] = PP(X_{t-\tau} \mid X_t) - PP(M_{t-\tau} \mid M_t) - PP(N_{t-\tau} \mid N_t)$$

Etc.

Gaussian case: Same as inf theory $\Phi$

Non-Gaussian case: pragmatic alternative

Enables comparison with (Granger) causal density
## Tononi approach vs Barrett - Seth approach

<table>
<thead>
<tr>
<th>Tononi</th>
<th>Barrett - Seth</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i) Identity with C.</td>
<td>(i) Correlation with C.</td>
</tr>
<tr>
<td>(ii) Fundamental functional units.</td>
<td>(ii) Choice of level e.g. brain regions.</td>
</tr>
<tr>
<td>(iii) Information about hypothetical past state in which all</td>
<td>(iii) Information about actual past state,</td>
</tr>
<tr>
<td>configurations equally likely.</td>
<td>based on current statistics.</td>
</tr>
<tr>
<td>(iv) Requires memoryless (Markovian) dynamics with finite set of</td>
<td>(iv) Not necessary to have memoryless</td>
</tr>
<tr>
<td>states. Inapplicable to neuroimaging data.</td>
<td>dynamics or finite set of states. Applicable</td>
</tr>
<tr>
<td>(v) State-dependent.</td>
<td>to neuroimaging data.</td>
</tr>
<tr>
<td>(vi) Compute from generative model.</td>
<td>(v) State-independent. Changes only with</td>
</tr>
<tr>
<td>(vii) C as combination of capacity and process.</td>
<td>changes to dynamics i.e. functional</td>
</tr>
<tr>
<td></td>
<td>connectivity.</td>
</tr>
<tr>
<td></td>
<td>(vi) Compute from statistics.</td>
</tr>
<tr>
<td></td>
<td>(vii) C as a process.</td>
</tr>
</tbody>
</table>

Several differences but in simulation very similar for simple systems.