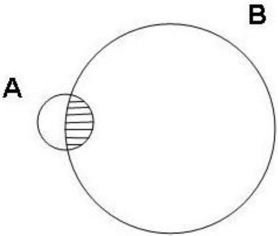
Bayesian Inference

Will Penny

24th February 2011

Bayes rule

Given probabilities $p(A)$, $p(B)$, and the joint probability $p(A,B)$, we can write the conditional probabilities



$$\begin{aligned} p(B|A) &= \frac{p(A,B)}{p(A)} \\ p(A|B) &= \frac{p(A,B)}{p(B)} \end{aligned}$$

Eliminating $p(A,B)$ gives Bayes rule

$$p(B|A) = \frac{p(A|B)p(B)}{p(A)}$$

Bayes rule

The terms in Bayes rule

$$p(B|A) = \frac{p(A|B)p(B)}{p(A)}$$

are referred to as the prior, $p(B)$, the likelihood, $p(A|B)$, and the posterior, $p(B|A)$.

The probability $p(A)$ is a normalisation term and can be found by *marginalisation*. For example,

$$\begin{aligned} p(A=1) &= \sum_B p(A=1,B) \\ &= p(A=1,B=0) + p(A=1,B=1) \\ &= p(A=1|B=0)p(B=0) + p(A=1|B=1)p(B=1) \end{aligned}$$

Medical Decision Making

Johnson et al (2001) consider Bayesian inference in for Magnetic Resonance Angiography (MRA). An Aneurysm is a localized, blood-filled balloon-like bulge in the wall of a blood vessel. They commonly occur in arteries at the base of the brain.

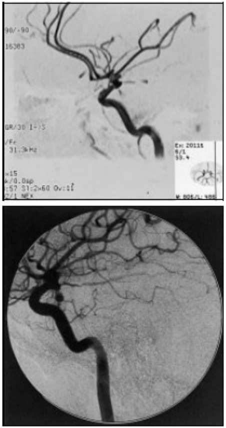


Fig 1 Case 1: magnetic resonance angiography with maximum intensity projection has normal appearance (top), whereas intra-arterial digital subtraction angiography (injection of left internal carotid artery) shows a large left posterior communicating artery aneurysm (bottom)

MRA can miss sizable Intracranial Aneurysms (IA)'s but is non-invasive (top).

Intra-Arterial Digital Subtraction Angiography (DSA) (bottom) is the gold standard method for detecting IA but is an *invasive* procedure requiring local injection of a contrast agent via a tube inserted into the relevant artery.

Medical Decision Making

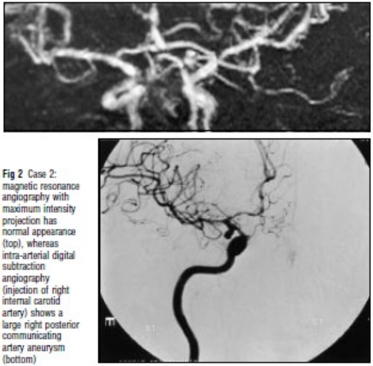


Fig 2 Case 2: magnetic resonance angiography with maximum intensity projection has normal appearance (top), whereas intra-arterial digital subtraction angiography (injection of right internal carotid artery) shows a large right posterior communicating artery aneurysm (bottom)

Given patient 1's symptoms (oculomotor palsy), the prior probability of IA (prior to MRA) is believed to be 90%.

For IAs bigger than 6mm MRA has a sensitivity and specificity of 95% and 92%.

What then is the probability of IA given a *negative* MRA test result ?

Medical Decision Making

The probability of IA given a negative test can be found from Bayes rule

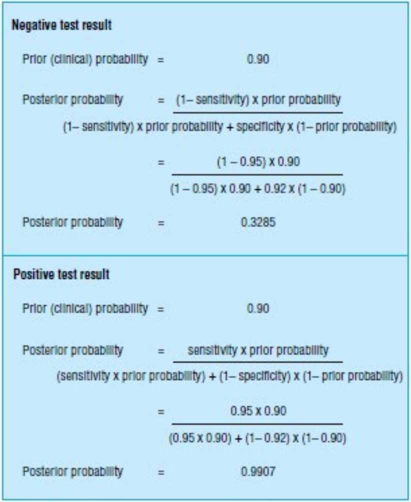
$$p(IA=1|MRA=0) = \frac{p(MRA=0|IA=1)p(IA=1)}{p(MRA=0|IA=1)p(IA=1)+p(MRA=0|IA=0)p(IA=0)}$$

where $p(IA=1)$ is the probability of IA prior to the MRA test. MRA test sensitivity and specificity are

$$p(MRA=1|IA=1)$$
$$p(MRA=0|IA=0)$$

We have $p(MRA=0|IA=1) = 1 - p(MRA=1|IA=1)$

Medical Decision Making



Negative test result

Prior (clinical) probability = 0.90

$$\text{Posterior probability} = \frac{(1-\text{sensitivity}) \times \text{prior probability}}{(1-\text{sensitivity}) \times \text{prior probability} + \text{specificity} \times (1-\text{prior probability})}$$

$$= \frac{(1-0.95) \times 0.90}{(1-0.95) \times 0.90 + 0.92 \times (1-0.90)}$$

Posterior probability = 0.3285

Positive test result

Prior (clinical) probability = 0.90

$$\text{Posterior probability} = \frac{\text{sensitivity} \times \text{prior probability}}{(\text{sensitivity} \times \text{prior probability}) + (1-\text{specificity}) \times (1-\text{prior probability})}$$

$$= \frac{0.95 \times 0.90}{(0.95 \times 0.90) + (1-0.92) \times (1-0.90)}$$

Posterior probability = 0.9907

Fig 3 Probability of a posterior communicating artery aneurysm given a negative or positive result from magnetic resonance angiography and a prior clinical probability of 90%. Sensitivity and specificity of angiography are 95% and 92% respectively. Probabilities are expressed between 0.0 (0%) and 1.0 (100%)

Medical Decision Making

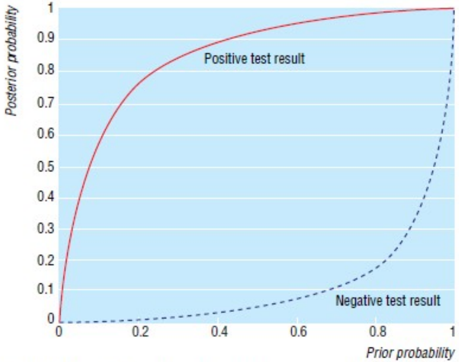
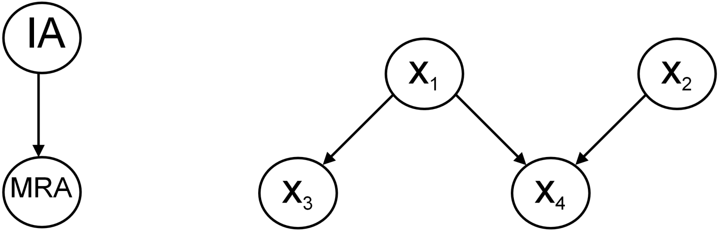


Fig 4 Influence of prior clinical probability on the probability of a disease after a negative or positive test result. Test sensitivity and specificity are 95% and 92% respectively

A negative MRA cannot therefore be used to exclude a diagnosis of IA. In both reported cases IA was initially excluded, until other symptoms developed or other tests also proved negative.

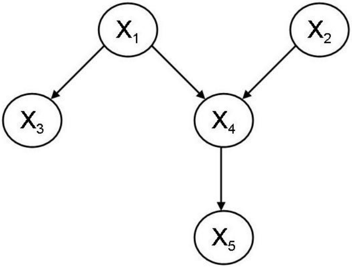
Multiple Causes and Observations

Multiple potential causes (eg. IA, X) and observations (eg. headache, oculomotor palsy, double vision, drooping eye lids, blood in CSF)



Directed Acyclic Graph

For a Directed Acyclic Graph (DAG)



The joint probability of all variables, x, can be written down as

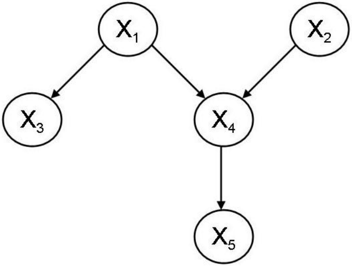
$$p(x) = \prod_{i=1}^{5} p(x_i | pa[x_i])$$

where $pa[x_i]$ are the parents of x_i.

Joint Probability

A DAG specfies the joint probability of all variables.

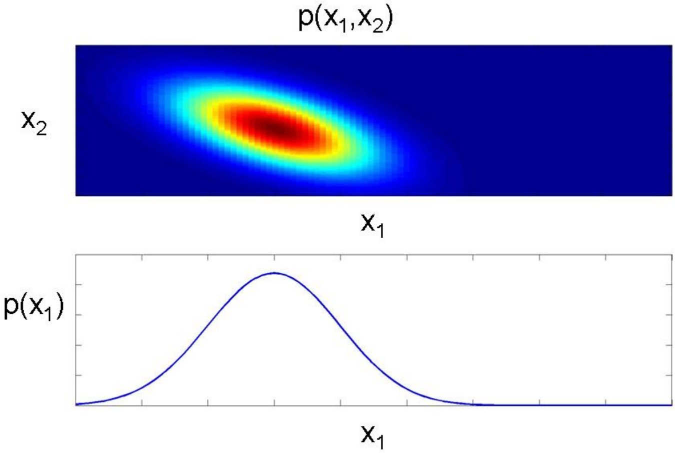
$$p(x_1, x_2, x_3, x_4, x_5) = p(x_1)p(x_2)p(x_3|x_1)p(x_4|x_1, x_2)p(x_5|x_4)$$



The negative log of the joint probability is known as the Gibbs Energy. All other variables can be gotten from the joint probability via marginalisation.

Marginalisation

$$p(x_1) = \int p(x_1, x_2) dx_2$$



Marginalisation

$$p(x_1, x_2) = \int \int \int p(x_1, x_2, x_3, x_4, x_5) dx_3 dx_4 dx_5$$

$$p(x_4) = \int \int \int \int p(x_1, x_2, x_3, x_4, x_5) dx_1 dx_2 dx_3 dx_5$$

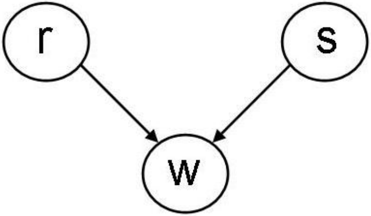
$$1 = \int \int \int \int \int p(x_1, x_2, x_3, x_4, x_5) dx_1 dx_2 dx_3 dx_4 dx_5$$

$$p(x_1) = \sum_{x_2} p(x_1, x_2)$$

$$p(x_2 = 3, x_3 = 4) = \sum_{x_1} p(x_1, x_2 = 3, x_3 = 4)$$

Did I Leave The Sprinkler On ?

A single observation with multiple potential causes (not mutually exclusive). Both rain, *r*, and the sprinkler, *s*, can cause my lawn to be wet, *w*.



$$p(w, r, s) = p(r)p(s)p(w|r, s)$$

Did I Leave The Sprinkler On ?

The probability that the sprinkler was on given i've seen the lawn is wet is given by Bayes rule

$$\begin{aligned} p(s=1|w=1) &= \frac{p(w=1|s=1)p(s=1)}{p(w=1)} \\ &= \frac{p(w=1,s=1)}{p(w=1,s=1)+p(w=1,s=0)} \end{aligned}$$

where the joint probabilities are obtained from marginalisation

$$\begin{aligned} p(w=1,s=1) &= \sum_{r=0}^{1} p(w=1,r,s=1) \\ p(w=1,s=0) &= \sum_{r=0}^{1} p(w=1,r,s=0) \end{aligned}$$

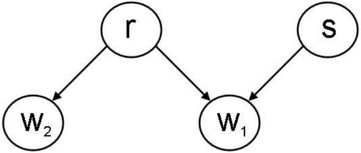
and from the generative model we have

$$p(w,r,s)=p(r)p(s)p(w|r,s)$$

Look next door

Rain r will make my lawn wet w_1 and nextdoors w_2 whereas the sprinkler s only affects mine.

$$p(w_1, w_2, r, s) = p(r)p(s)p(w_1|r, s)p(w_2|r)$$



After looking next door

Use Bayes rule again

$$p(s=1|w_1=1,w_2=1) = \frac{p(w_1=1,w_2=1,s=1)}{p(w_1=1,w_2=1,s=1)+p(w_1=1,w_2=1,s=0)}$$

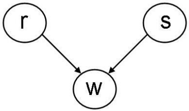
with joint probabilities from marginalisation

$$p(w_1=1,w_2=1,s=1) = \sum_{r=0}^{1} p(w_1=1,w_2=1,r,s=1)$$

$$p(w_1=1,w_2=1,s=0) = \sum_{r=0}^{1} p(w_1=1,w_2=1,r,s=0)$$

Numerical Example

Bayesian models force us to be explicit about exactly what it is we believe.

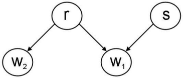


$$
\begin{aligned}
p(r=1) &= 0.01 \\
p(s=1) &= 0.02 \\
p(w=1|r=0,s=0) &= 0.001 \\
p(w=1|r=0,s=1) &= 0.97 \\
p(w=1|r=1,s=0) &= 0.90 \\
p(w=1|r=1,s=1) &= 0.99
\end{aligned}
$$

These numbers give

$$
\begin{aligned}
p(s=1|w=1) &= 0.67 \\
p(r=1|w=1) &= 0.31
\end{aligned}
$$

Explaining Away



Numbers same as before. In addition

$$p(w_2 = 1 | r = 1) = 0.90$$

Now we have

$$p(s = 1 | w_1 = 1, w_2 = 1) = 0.21$$
$$p(r = 1 | w_1 = 1, w_2 = 1) = 0.80$$

The fact that my grass is wet has been explained away by the rain (and the observation of my neighbours wet lawn).

Perception as Inference

In Helmholtz's view our percepts are our best guess as to what is in the world, given both sensory data and prior experience. He proposed that perception is unconscious inference.

Gaussians

Precision is inverse variance eg. a variance of 0.1 is a precision of 10.

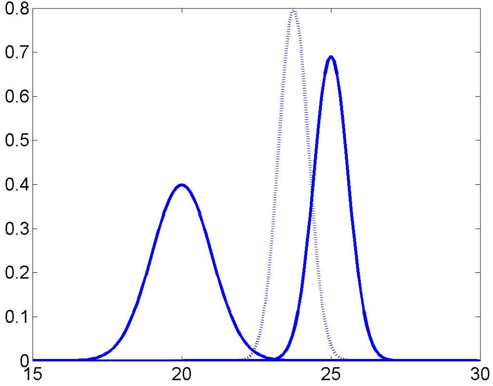
For a Gaussian prior with mean m_0 and precision λ_0, and a Gaussian likelihood with mean m_D and precision λ_D the posterior is Gaussian with

$$\begin{aligned} \lambda &= \lambda_0 + \lambda_D \\ m &= \frac{\lambda_0}{\lambda} m_0 + \frac{\lambda_D}{\lambda} m_D \end{aligned}$$

So, (1) precisions add and (2) the posterior mean is the sum of the prior and data means, but each weighted by their relative precision.

Gaussians

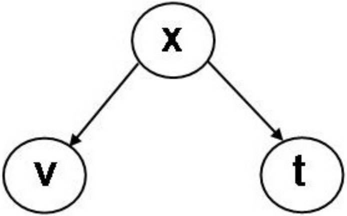
The two solid curves show the probability densities for the prior $m_0 = 20$, $\lambda_0 = 1$ and the likelihood $m_D = 25$ and $\lambda_D = 3$. The dotted curve shows the posterior distribution with $m = 23.75$ and $\lambda = 4$. The posterior is closer to the likelihood because the likelihood has higher precision.



Sensory Integration

Ernst and Banks (2002) asked subjects which of two sequentially presented blocks was the taller. Subjects used either vision alone, touch alone or a combination of the two.

If vision v and touch t information are independent given an object x then we have



$$p(v,t,x) = p(v|x)p(t|x)p(x)$$

Bayesian fusion of sensory information then produces a posterior density

$$p(x|v,t) = \frac{p(v|x)p(t|x)p(x)}{p(v,t)}$$

Sensory Integration

In the abscence of prior information about block size (ie $p(x)$ is uniform), for Gaussian likelihoods, the posterior will also be a Gaussian with precision λ_{vt}. From Bayes rule for Gaussians we know that precisions add

$$\lambda_{vt} = \lambda_v + \lambda_t$$

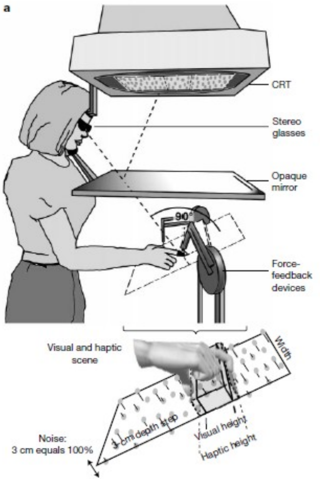
and the posterior mean is a relative-precision weighted combination

$$\begin{aligned} m_{vt} &= \frac{\lambda_v}{\lambda_{vt}} m_v + \frac{\lambda_t}{\lambda_{vt}} m_t \\ m_{vt} &= w_v m_v + w_t m_t \end{aligned}$$

with weights w_v and w_t.

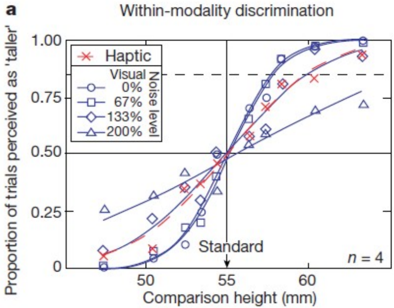
Vision and Touch

Ernst and Banks (2002) asked subjects which of two sequentially presented blocks was the taller. Subjects used either vision alone, touch alone or a combination of the two.



Vision and Touch Separately

They recorded the accuracy with which discrimination could be made and plotted this as a function of difference in block height. This was first done for each condition alone. One can then estimate precisions, λ_v and λ_t by fitting a cumulative Gaussian density function.

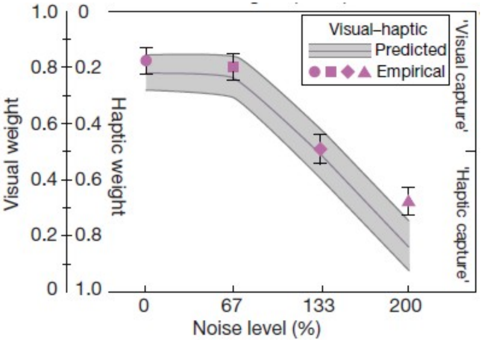


They manipulated the accuracy of the visual discrimination by adding noise onto one of the stereo images.

Vision and Touch Together

Optimal fusion predicts weights from Bayes rule

$$
\begin{aligned}
\lambda_{vt} &= \lambda_v + \lambda_t \\
m_{vt} &= \frac{\lambda_v}{\lambda_{vt}} m_v + \frac{\lambda_t}{\lambda_{vt}} m_t \\
m_{vt} &= w_v m_v + w_t m_t
\end{aligned}
$$



They observed visual capture at low levels of visual noise and haptic capture at high levels.

Decision Making Dynamics

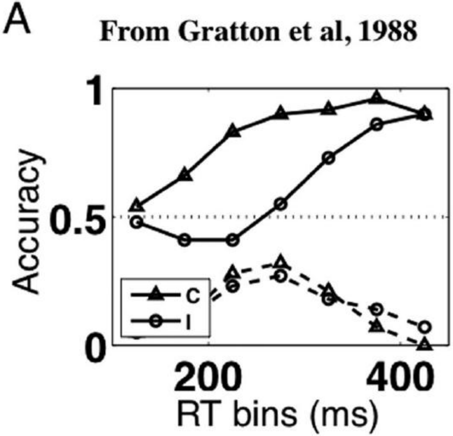
In the Eriksen Flanker task subjects have to implement the following stimulus-response mappings

Stimulus	*Response*
1.*HHH*	*Right*
2.*SHS*	*Right*
3.*SSS*	*Left*
4.*HSH*	*Left*

Put simply, the subject should press the right button if the central cue is H and left if it is S. On trial type one and three the flankers are compatible ($M = C$) and on two and four they are incompatible ($M = I$).

Decision Making Dynamics

If subjects are too slow an auditory beep is emitted. This is the *deadlined* Flanker task.

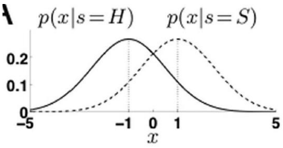


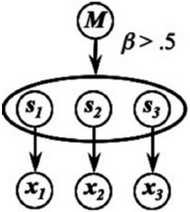
On incompatible trials initial average accuracy dips below the chance level.

Likelihood

Yu et al. (2009) assume three populations of neurons, x, that are driven by the three stimuli, s

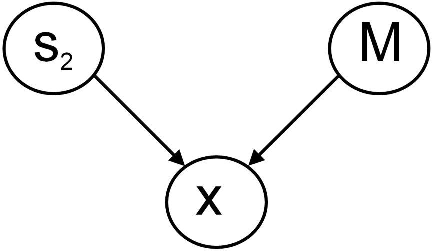
$$p(x|s) = \prod_{i=1}^{3} \mathrm{N}(x_i; \mu_i, \sigma^2)$$





$$\begin{aligned} p(x|s=SHS) &= p(x|s_2=H, M=l) \\ &= \mathrm{N}(x_1; 1, \sigma^2)\mathrm{N}(x_2; -1, \sigma^2)\mathrm{N}(x_3; 1, \sigma^2) \end{aligned}$$

Generative Model



Joint probability

$$p(x, s_2, M) = p(x|s_2, M)p(s_2)p(M)$$

Likelihood

$$p(x|s_2, M) = \prod_{i=1}^{3} p(x_i|s_2, M)$$

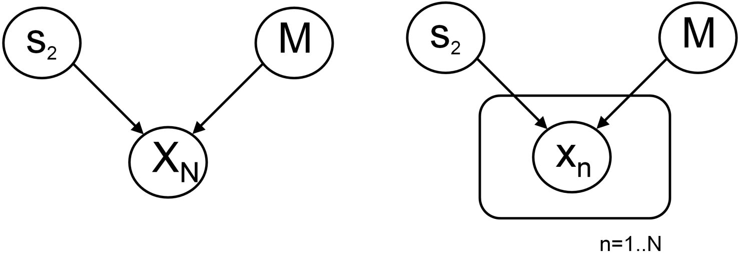
Dynamics

Consider a discrete set of time points $t(n)$ within the trial with $n = 1, 2, ..N$.

Denote x_n as population vector observed at time $t(n)$ and $X_n = [x_0, x_1, ..., x_n]$ as all vectors observed up until time point $t(n)$.

Yu et al. (2009) formulate a discrete time inferential model. We will consider continuous time models later.

Generative Model



Joint probability

$$p(X_N, s_2, M) = p(X_N|s_2, M)p(s_2)p(M)$$

Likelihood

$$p(X_N|s_2, M) = \prod_{n=1}^{N} p(x_n|s_2, M)$$

Inference

The following joint probability is updated recursively

$$p(s_2, M|X_n) = \frac{p(x_n|s_2, M)p(s_2, M|X_{n-1})}{\sum_{s_2', M'} p(x_n|s_2', M')p(s_2', M'|X_{n-1})}$$

Then marginalise over M to get decision probability

$$p(s_2 = H|X_n) = p(s_2 = H, M = C|X_n) + p(s_2 = H, M = I|X_n)$$

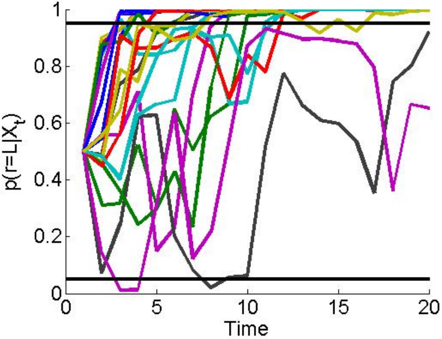
Initialise with

$$\begin{aligned} p(s_2 = H, M = C|X_0) &= p(s_2 = H)p(M = C) \\ p(s_2 = H, M = C|X_0) &= 0.5\beta \\ p(s_2 = H, M = I|X_0) &= 0.5(1 - \beta) \end{aligned}$$

where $p(M = C) = \beta$.

Inference

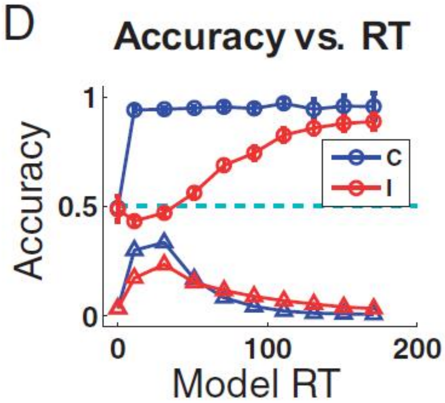
On most trials (18 out of 20) evidence slowly accumulates in favour of the central stimulus being $s_2 = H$. This is reflected in the posterior probability $p(s_2 = H|X_n)$.



This corresponds to evidence for a left button press.

Compatibility Bias Model

For compatibility bias $\beta > 0.5$



The model also shows the initial dip for incompatible flankers.

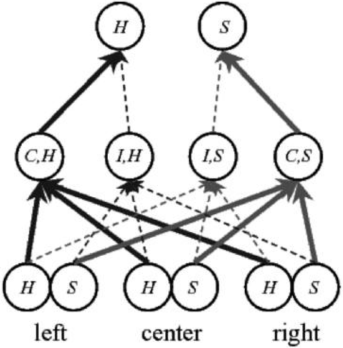
Neural Implementation

The Bayesian inference equations

$$p(s_2, M|X_n) = \frac{p(x_n|s_2, M)p(s_2, M|X_{n-1})}{\sum_{s_2', M'} p(x_n|s_2', M')p(s_2', M'|X_{n-1})}$$

$$p(s_2 = H|X_n) = p(s_2 = H, M = C|X_n) + p(s_2 = H, M = I|X_n)$$

can be implemented as a network model.



The hidden layer representations are *self-exciting* and require *divisive normalisation*. In the compatibility bias model the compatible pathway is initially excited.

Approximate Inference

As the number of stimuli grows exact inference becomes intractable. Instead, we can initially *assume* compatibility.

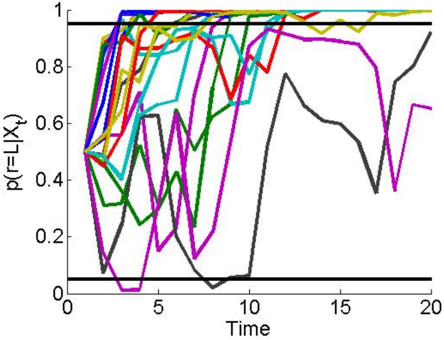
$$p(s_2 = H|X_t) = \frac{p(x_1(t)|s_1 = H)p(x_2(t)|s_2 = H)p(x_3(t)|s_3 = H)p(s_2 = H|X_{t-1})}{\sum_{s=H,S} p(x_1(t)|s_1 = s)p(x_2(t)|s_2 = s)p(x_3(t)|s_3 = s)p(s_2 = s|X_{t-1})}$$

If the flankers are detected to be incompatible we can then switch to an inferential scheme which ignores them

$$p(s_2 = H|X_t) = p(x_2(t)|s_2 = H)p(s_2 = H|X_{t-1})$$

Conflict detection

Compatibility can be inferred from a conflict detector

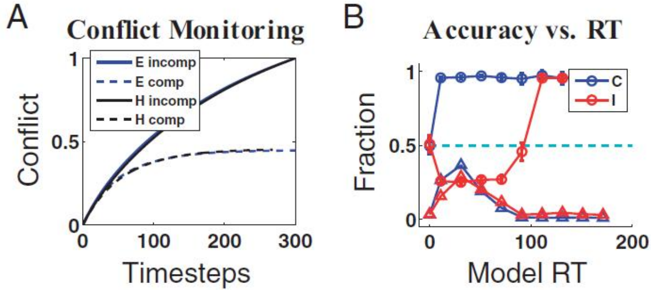


which measures the energy in the decision region (Botvinick et al. 2001)

$$E_t = E_{t-1} + p(s_2 = H|X_t)p(s_2 = S|X_t)$$

Approximate Inference

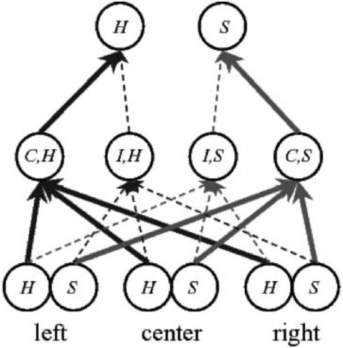
Detecting conflict using an energy measure gives similar results to using an entropy measure, *H*



Approximate inference yields behaviour similar to exact inference and empirical data.

Neural Implementation

Output of conflict monitoring enhances $M = C$ or $M = I$ pathway.



References

C. Bishop (2006) Pattern Recognition and Machine Learning, Springer.

M. Botvinick et al. (2001) Psych Review 108, 624-652.

M. Ernst and M. Banks (2002) Nature 415, 429-433.

M. Johnson et al. (2001) BMJ 322, 1347-1349.

D. Mackay (2003) Information Theory, Inference and Learning Algorithms, Cambridge.

D. Wolpert and Z. Ghahramani (2004) In Gregory RL (ed) Oxford Companion to the Mind, OUP.

A. Yu, P. Dayan and J. Cohen (2009) J Exp Psych 35,700-717.