Bayesian models in cognitive neuroscience: A tutorial

Jill X. O'Reilly & Rogier B. Mars

To appear in: *An introduction to model-based cognitive neuroscience* (Forstmann BU, Wagenmakers EJ, Eds.) Springer (in press)

Jill X. O'Reilly
Centre for Functional MRI of the Brain (FMRIB), Nuffield Department of Clinical Neurosciences, John Radcliffe Hospital, University of Oxford, Oxford OX3 9DU, United Kingdom, E-mail: joreilly@fmrib.ox.ac.uk

Rogier B. Mars
Department of Experimental Psychology, University of Oxford, South Parks Road, Oxford OX1 3UD, United Kingdom, E-mail: rogier.mars@psy.ox.ac.uk

Abstract

This chapter provides an introduction to Bayesian models and their application in cognitive neuroscience. The central feature of Bayesian models, as opposed to other classes of models, is that Bayesian models represent the beliefs of an observer as probability distributions, allowing them to integrate information while taking its uncertainty into account. In the chapter, we will consider how the probabilistic nature of Bayesian models makes them particularly useful in cognitive neuroscience. We will consider two types of tasks in which we believe a Bayesian approach is useful: optimal integration of evidence from different sources, and the development of beliefs about the environment given limited information (such as during learning). We will develop some detailed examples of Bayesian models to give the reader a taste of how the models are constructed and what insights they may be able to offer about participants' behavior and brain activity.

Introduction

In the second half of the 18th century, the French mathematician Pierre-Simon Laplace was confronting a dilemma. He wanted to use observations of the location of planets to test the predictions made recently by Isaac Newton about the motion of heavenly bodies and the stability of the solar system. However, the data Laplace was confronted with was cobbled together from sources all over the word and some of it was centuries old. Couple that with the imprecision of the instruments of the time and Laplace had what we now call noisy data on his hands.

Laplace decided he needed a method that would allow him to use the large amounts of data obtained by astronomers, some of which might be unreliable, to determine the real state of the universe they were observing. In other words, he needed a way to move back from observed events (the astronomers' observations) to the most probable cause (the position of a planet). In doing so he created a way of thinking fundamentally different from the established approaches at the time. Because Laplace unwittingly hit upon elements of an earlier work by the English reverent Thomas Bayes [1], we now know this way of thinking as 'Bayesian'.

The Bayesian way of thinking is so different from conventional statistics that it was 'non grata' in most university departments for a long time. Only since the mid-20th century has this begun to change. Bayesian methods were starting to be applied pragmatically to solve a host of real-world problems. Moreover, the invention of computers enabled people to perform the often labor intensive computations required in Bayesian statistics automatically. Slowly, Bayesians dared to come out of the closet (see McGrayne [2] for a history of Bayesian thinking). Bayesian thinking has now been applied to almost every field imaginable, including code-breaking, weather prediction, improving the safety of coal mines, and—most relevant for the purpose of this book—the modeling of human behavior and human brain function.

This chapter is about the use of Bayesian models in cognitive neuroscience and psychology, with particular reference to the modeling of beliefs and behavior. The reason for using formal models in this context is to gain insight into internal

representations of the environment and of experimental tasks that are held by participants, and to use them to predict behavior and brain activity. We will therefore begin by explaining how the representation of the world contained in a Bayesian model (or brain) differs from non-Bayesian representations, and go on to consider how these features can be used in the context of cognitive neuroscience and psychology research.

We will first discuss three key features of Bayesian system: Bayesian systems represent beliefs as probability distributions, Bayesian systems weight different sources of information according to their associated uncertainty, and Bayesian systems interpret new observations in the light of prior knowledge.

After considering how a Bayesian model's worldview differs from that of a non-Bayesian model, we will briefly review some evidence from the psychology and neuroscience literature that human and animal observers behave in a Bayesian manner. In particular we will focus on two classes of problems in which Bayesian models behave differently from non-Bayesian ones: integration of sensory evidence from different sources, and learning.

In the final section of the chapter, we will look in more detail at how Bayesian models can be constructed and what insights can be gained from them. We will consider Bayesian approaches to two problems: inferring a spatial distribution from a few observations, and inferring the probability of targets or rewards appearing in one of two locations in a gambling task. By constructing Bayesian 'computer participants' for each of these tasks, we will gain insights into factors that might predict the performance of human or animal participants on the same tasks.

The defining features of a Bayesian model

Bayesian statistics is a framework for making inferences about the underlying state of the world, based on observations and prior beliefs. The Bayesian approach, to try and infer *causes* from their observed *effects,* differs philosophically from other approaches

to data analysis. Other approaches, often referred to as 'frequentist' approaches, focus on obtaining summary statistics for the observed data (such as the mean or expected value of an observation) without reference to the underlying causes that generated the data.

Bayesian systems represent beliefs as probability distributions

A Bayesian approach to understanding data is to consider a range of possible causes for the observed data, and assign probabilities to each of them. A subtle but crucial consequence of this approach is that, although the true state of the environment takes a single value, the observer's idea of the environment can be represented as a continuous distribution over many possible states of the environment. In other words, even though the observer knows there is only one true cause of his observations, he can still assign a graded probability to several possible causes. The observer's model represents these possibilities as a *probability density function* (pdf). This single feature, the representation of beliefs as probability density functions, gives rise to much of the behavior that differentiates Bayesian models from non-Bayesian ones.

Let's illustrate the use of probability density functions with an example. Consider the following scenario: Isaac Newton is foraging for apples in his garden when he sees an apple fall from a tree into long grass (Fig. 1a). Where should he go to retrieve the apple? If he saw the apple fall into the undergrowth, then the most likely place to look for the apple is near where it fell. We might, therefore, represent his belief about the location of the apple (his *internal model* of the state of the environment) as a single value, the location at which the apple entered the undergrowth (Fig. 1b; for simplicity, let's assume we can represent the location in a one-dimensional space). However, because the apple is now out of sight, Isaac can't be certain exactly where it is (it may have rolled along the ground in an unknown direction). This uncertainty can be incorporated into his internal model, if instead of using a single value the apple's position is represented as a probability distribution. Then we can make statements like 'there is a 95% chance that there will be an apple between locations A and B' (Fig. 1c). Note that as well as the most likely location of the apple (the model of

the distribution) this representation captures uncertainty (the width or variance of the distribution).

<Please insert Figure 1 about here>

Note that the Bayesian use of probability density functions to represent degree of belief about a single state of the world is rather distinct from the typical use of probability density functions to represent the frequency of observations. In our apple example, Isaac Newton knows a *single* apple fell from the tree, and represents the location of that *single* apple as a probability density function, although in fact there is only one apple and it has only one true location. A more typical (frequentist) construction of a probability density function would be to represent the *frequency* with which apples were observed in different locations. Whilst for a hundred apples, the frequentist and Bayesian pdfs may look the same, for a single apple, the frequentist view is that the apple is either in a position, or it is not.

Bayesian systems integrate information using uncertainty

When a belief about the state of the world (for example, about the location of an apple) is represented as a probability density function, the variance of that pdf, in other words the width of the pdf, represents the degree of uncertainty about the state of the world. One key feature of the Bayesian approach is that Bayesian systems take this uncertainty into account and use it to weight different sources of information according to their relative precisions.

Imagine that Isaac Newton has both *seen* an apple fall from the tree into long grass, and *heard* it hit the ground. His belief about the location of the apple based on each of these sources of information (vision and hearing) can be represented as a single probability density function. How should Isaac's brain use these two sources of information to get the best estimate of the apple's location? One solution is to use only the more reliable, or preferred sense. But this wastes the information from the other sense. A better solution is to combine the estimates of location based on vision and hearing.

How should the two sensory modalities be combined? Perhaps Isaac could take a point midway between the most likely location given what he saw, and the most likely location given what he heard? The Bayesian solution to this problem is to apply *precision weighting* the two sources of information, that is to give more weight to the observation with the lowest variance. If, for example, vision gives a more precise estimate of where the apple fell, then visual evidence should be given more weight. On the other hand, if vision is unreliable (e.g. at night), auditory evidence should be given more weight.

Let's look at this graphically. In figure 2a, we can see that the pdf of Isaac's visual information (in red) is much less wide than his pdf based on hearing (in blue). Or, to put it more precisely, the variance of the vision pdf is smaller than that of the hearing pdf. Thus, optimally combining these two sources of information will result in a pdf closer to the vision pdf. However, in figure 2b, the vision is much less reliable, indicated by a greater variance in the red pdf. The combined pdf now is much closer to the hearing one.

<Please insert Figure 2 about here>

Precision weighting is only possible if the observations (by vision and hearing) are represented as probability density functions; if each observation was represented in terms of a single most likely location, we could still combine the predictions by taking a point between the locations given vision and hearing, but there would be no way to take into account the relative reliability or precision of the two sources of information. However, given that observations are represented as probability density functions, precision weighting arises naturally from simply multiplying together the two probability distributions[1]. Then the probability of the apple's location given both visual and auditory information is highest where the two distributions overlap, and

[1] In fact, the probability of each location given hearing and vision can only be obtained by multiplication if the variance in the two probability density functions is independent. In this case, we are talking about uncertainty that arises from noise in the sensory systems, which we can safely assume is independent between vision and hearing.

the mode (peak) of the combined distribution lies closer to the mode of the distribution with the lowest variance.

Bayesian systems interpret new information in the light of prior knowledge

Isaac Newton probably had some previous experience with apples falling from trees. Therefore, it would seem sensible if he used this prior knowledge to inform his model of where the apple might lie. For example, he might have some expectations about how far the apple might roll, the slope of the land, etc. Even if Isaac didn't see an apple fall, he would still have a prior belief that apples should be found under the apple tree – not, for example, under the lamppost. Isaac knows the apple should not fall far from the tree.

In the same way the location of the apple, given Isaac saw it fall, can be represented as a probability density function, so can his prior beliefs. In Bayesian thinking these prior beliefs are called the *priors*. Furthermore, current observations can be combined with the prior, just as probability density functions based on vision and hearing were combined in the previous section. Combining the current observations with the prior gives a *posterior* distribution that takes both into account.

The ability to combine current observations with a prior, or to combine parallel sources of information like vision or hearing, is embodied in the central theorem of Bayesian statistics, called *Bayes' theorem*:

$$p(\text{true state} \mid \text{observation}) \propto p(\text{observation} \mid \text{true state}) \times p(\text{true state})$$

Eq. 1

… where p(true state) is defined as the probability that a given hypothetical state of the environment (such as a given location for a planet or an apple) was true, based on all sources of information other than the observation currently being considered. The term 'other sources of information' can equally well include other sensory modalities or prior knowledge.

In Bayesian terminology, the left hand side of Equation 1, p(true state | observation), is called the *posterior*; the expression p(observation | true state) is called the *likelihood*; and p(true state) is called the *prior*. Bayes' theorem thus says that our belief about the true state of the environment after our observations is proportional to our prior beliefs weighted by the current evidence.

Priors and learning

Because Bayes' theorem tells us how we should combine new observations with prior beliefs, it provides particularly useful insights about how the observer's beliefs should evolve in situations where information about the environment is obtained sequentially. For example, we can model how Isaac's beliefs evolve while he observes a number of falling apples. After each observation, he updates his prior to a new posterior. This posterior then serves as the new prior for the next apple.

In experimental paradigms in which participants learn by trial and error, we cannot assume the observer has complete knowledge of the state of the environment. These paradigms are a key target for model-based cognitive neuroscience, since if we want to model a participant's behavior or brain activity, it is arguably more appropriate to base our predictions on a model of what the participant might *believe* the state of the environment to be, rather than basing our predictions about brain activity on the true state of the environment, which the participant could not in fact know, unless he/she/it was clairvoyant.

Of course, not all learning models are Bayesian – for example, temporal-difference learning models such as the Rescorla-Wagner algorithm are also popular. Non-Bayesian class of algorithms can do a good job of explaining behavior in many experiments. In a later section of the tutorial we will investigate the differences between Bayesian and non-Bayesian learning algorithms in more detail, in order to highlight cases where Bayesian models can give us enhanced insights into the participant's thought processes as compared to non-Bayesian learning algorithms.

Are Bayesian models valid for modeling behavior?

In the previous section we've seen how Bayesian thinking can be used to model the beliefs of an observer and can track how these beliefs should evolve when combining different sources of information or during learning based on repeated observations. Mathematically, it can be shown that the Bayesian approach is the best approach to combine information under uncertainty with the greatest precision [3]. However, for these models to be useful in cognitive neuroscience we need to know if people combine information in similar ways. Fortunately, it turns out they often do. People and animals can show behavior close to the optimum predicted by Bayesian theory. In this section, we will provide some examples of how human behavior can be described by Bayesian models. We will limit ourselves to illustrating how human behavior shows some of the Bayesian characteristics we described above. More in-depth reviews of how the Bayesian approach can inform our understanding of behavior and brain function are provided by O'Reilly [4], Chater and Oaksford [5], and Körding and Wolpert [6].

At the most fundamental level, one can see the human brain as a device whose job it is, at least partly, to infer the state of the world. However, we know that the nervous system is noisy. Thus we need to deal with information under uncertainty. One way that psychologists have suggested we deal with this is the use of 'top-down information'. In the terms of Bayesian theory this means people have a prior that influences their information processing. The effect of such a prior has been demonstrated in vision by the existence of a variety of well-known visual illusions. For instance, in the famous Müller-Leyer illusion people see two line segments of equal length that have short lines on their ends, either pointing in the direction of the line or away from it. Most people report the second line to be longer than the first. Gregory [7] suggested this is because people have priors about perspective that they have learned from the buildings in the environment, in which the former configuration corresponds to an object which is closer and the latter with an object far away. Interestingly, this predicts that people who have grown up in a different environment might not have this illusion. This indeed seems to be the case for some African tribes [8].

In our daily life, we often have to reconcile different sources of information. As shown above, Bayesian thinking implies that different sources of information should be combined using precision weighting. As one illustration of whether humans combine information in this way, Jacobs [9] asked participants to match the height of an ellipse to the depths of a simulated cylinder defined by texture and motion cues. The participants were either given motion information, texture information, or both about the depth of the cylinder. Bayesian models predicted how participants combined the sources of information. Similarly, Ernst and Banks [10] asked participants to combine visual and haptic information to judge the height of a raised bar. Participants were given conflicting information with the experimenter manipulating the precision of the information available by introducing noise in the visual stimulus. They reported that participants took the reliability of the visual information into account when combining the visual and haptic information, in a way that was predicted by Bayes' theorem.

Once we are satisfied that humans are able to behave in a Bayes' optimal fashion in general it becomes interesting to see in which situations their optimality breaks down. O'Reilly [4] discusses some instances in which a deviation from Bayesian predictions informs us about the limits of our cognitive system.

The usefulness of Bayes' theorem for modeling behavior and its particular characteristics are perhaps best illustrated during learning. Therefore we will spend the remainder of this chapter looking at the behavior of Bayesian systems during the learning of environmental contingencies.

Learning

As experimenters in cognitive neuroscience, we create the experimental environment in which our participants produce behavior. Therefore, we know the true parameters of the environment (in the previous example, this would be equivalent to knowing

where Newton's apple actually is). However, the participant does not know these true values; s/he must infer them from observations.

Since we are interested in the behavior and brain activity of the participant, it is advantageous to have an estimate of what the participant knows or believes about the state of the environment, as this might differ from the true state. This is particularly true when data about the environment are presented sequentially, as in many psychological tasks. For example, in the gambling tasks such as the one-armed, two-armed and multi-armed bandit tasks, participants, from humans [11, 12] to the humble bumble bee [13], learn the probability of rewards associated with certain actions by trial and error; similarly in uncued attentional tasks such as the uncued Posner task [14], participants learn over many trials that targets are more likely to appear in certain locations than others. In these sequential tasks, a number of trials must be experienced before the participant's estimates of the probabilities associated with each action approach the true values; in environments that change, continuous learning may be required.

'Today's posterior is tomorrow's prior'

As we briefly suggested above, learning from a sequence of observations can be modeled using iterative application of Bayes' rule. For example, let's say we observe a number of apples falling to the ground at locations x_1, x_2... x_i, and we want to infer from this the most likely location of fallen apples. Let's make the assumption that the distribution of apples is Gaussian about the tree trunk, with unknown mean μ (the location of the tree trunk) and variance σ^2. Then we can say that the variable x, the location of any given apple, follows a Gaussian distribution $x \sim \mathcal{N}(\mu, \sigma^2)$. Our aim is to infer the values of μ and σ^2 from the observed values of x.

Let's assume we have no a-priori knowledge about where the apple might fall. In other words, we start with a prior distribution such that all possible values of the parameters μ, σ^2 are considered equally likely. This is called a *flat prior*.

Then, on trial 1, we observe a data point, say x_1 = 67. Remember that Bayes' rule (Equation 1) tells us we can update our prior (which is flat) by our likelihood to give our posterior. In our current situation, we can determine the likelihood, since we know for each possible pair of parameters μ, σ^2 the probability that a value of 67 would have been observed. In this case this is the probability density of a Gaussian $\mathcal{N}(\mu, \sigma^2)$ for a region about the value 67 with unit width. Thus, we can work out the probability of each possible pair of parameters μ, σ^2 given this one observation, and plot a probability density function over 'parameter space' – the range of possible values of μ and σ^2 (Fig. 3a). This probability density distribution based on the current observation is sometimes called the 'likelihood function'.

<Please insert Figure 3 about here>

To obtain the posterior probability for each pair of values μ, σ^2 (the left hand side of Bayes' rule), we also need to take into account the prior probability that the values μ, σ^2 are correct by multiplying the likelihood function with the prior probability distribution. On trial one, we had a uniform prior, so the posterior is equal to the likelihood distribution. On trial two, we use the posterior resulting from trial one as be basis for our new prior. Again we observe a data point and update our prior to a new posterior that functions as the prior on the new trial. Etcetera. In general, what happens during learning is that the prior at trial *i* is derived from the posterior at trial *i-1*. Hence we can write Bayes' rule on trial *i* as follows:

$$p(x \sim \mathcal{N}(\mu, \sigma^2) | \, x_{1:i}) \propto p(x_i | \, x \sim \mathcal{N}(\mu, \sigma^2)) \times p(x \sim \mathcal{N}(\mu, \sigma^2) | x_{1:i-1}) \quad \text{Eq 2}$$

Thus, the posterior distribution on trial *i* is proportional to the likelihood of the observed data, x_i, times the prior distribution at trial *i*, which was derived from the posterior at trial *i-1* and captures all that is known about how previous data $x_{1:i-1}$ predict the current data point x_i.

In the mind of our Bayesian participant

We can now look into the 'mind' or our Bayesian computer participant to see what it knows and believes about the environment on a trial-to-trial basis. After the first observation, the posterior distribution (our model's estimate of the true values of μ, σ^2) has a peak at $\mu = 67$ and a very low value for σ^2, since all observed apples (all one of them) fell near to *x*=67.

We can represent the posterior over μ, σ^2 graphically on a grid (Fig. 3a, left panel) that represents all the possible combinations of mean and variance and their associated probabilities, which are denoted by colour. The space of all possible values for μ, σ^2 is called the *state space* or *parameter space*.

The next data point is x_2 = 100. This point is far from the previous observation. Therefore, the model's estimates shift. The estimated μ moves towards a point between 67 and 100, and the estimated σ^2 increases to create a distribution that encompasses *both* data points. As you can see the best fit Gaussian is now a much wider distribution, with a mean to somewhere in between 67 and 100 and a variance such that both data points are encompassed (Fig. 3b, left panel). However, the model is also relatively uncertain about the values of μ, σ^2 as can be seen from the wide spread of probability density across parameter space. As the model obtains more and more data the posterior distribution converges on a mean and standard deviation, and uncertainty decreases (Fig. 3c-d, left panels).

We can translate our model's estimates of the values μ, σ^2 into a probability density function over physical space (i.e., plot probability as a function of the possible positions, *x,* at which apples could fall). We do this in the right-hand panels of Figure 3.

To plot probability density over space, we need to decide how to summarize the distribution over parameter space, i.e. over μ, σ^2, which in fact represents our degree of belief in a range of different Gaussian distributions with different values of μ, σ^2. How should the distribution in parameter space be translated into a distribution over x? One option is to take the peak (or mode) of the distribution over μ, σ^2 – the values at the deepest red spot in the left-hand panels of Figure 3. This gives the most likely

Gaussian distribution (the maximum likelihood estimate). The resulting distributions are shown in black/red in the right hand panels of Figure 3. However, this measure ignores uncertainty about that distribution, throwing away a lot of information. Another option is to take a weighted sum of all possible Gaussian distributions over space – as given by[2]:

$$p(x) = \sum_{\mu} \sum_{\sigma^2} p(x|x \sim \mathcal{N}(\mu, \sigma^2)) \times p(x \sim \mathcal{N}(\mu, \sigma^2)|x_{1:i})$$

Eq. 3

This gives a distribution over x that takes into account variance due to uncertainty over μ, σ^2. The resulting distributions are shown in grey/blue in the right hand panels of Figure 3.

What is it useful for?

Using a Bayesian learner that iteratively integrates observed data with what is known from previous observations allows us to follow the dynamics of the different model parameters on a trial-by-trial basis. Using this trial-by-trial information, we can make predictions about behaviour – i.e. where Isaac should forage for apples on each trial. For example, we might hypothesize that he will search in an area centered on the estimated value of μ (i.e., he will search around where he thinks the tree is) and that the size of the area he searches in should be proportional to σ^2.

Furthermore, because the Bayesian model represents Isaac's beliefs about μ and σ^2 as probability distributions, our Bayesian model gives us an estimate of how uncertain he should be about those values (the spread of probability density in parameter space). Because we have this insight into uncertainty, which is the defining feature of

[2] In all the examples and exercises given here, we obtain an approximate solution by evaluating *p(x)* for discrete values of (μ, σ^2). In the continuous case, equation 3 would become:

$$p(x) = \int d\mu \int d\sigma^2 \, [p(x|x \sim \mathcal{N}(\mu, \sigma^2)) \times p(x \sim \mathcal{N}(\mu, \sigma^2)|x_{1:i})]$$

Bayesian models, we can make and test predictions about behavior based on uncertainty about the environment (estimation uncertainty [15, 16]) – for example, we might expect Isaac to express more exploratory behavior when uncertainty about μ and σ^2 is high [17, 18]. Moreover, we might expect that certain parameters of our Bayesian model might be reflected in neural activity. Although one has to be careful with interpretation [19], it is possible to link the values of the model's parameters to brain activity.

Another example of a Bayesian learner: one-armed bandit

In the previous example, we showed how a Bayesian computer participant could be used to model what a human participant knows or believes about the parameters of a Gaussian distribution from which spatial samples (the location of apples) were drawn. In fact, this is one example in which the beliefs of the participant (at least, an optimal Bayesian participant) rapidly approach the true state of the environment as can be seen in Figure 3.

There are many tasks, including some in common use in cognitive neuroscience, where an internal model based on sampling of the environment is a much weaker approximation of the true state of the environment. One such example is given by tasks in which the environment changes frequently (so the observer must constantly update his model of the environment) [11]. Another case is presented by tasks in which the parameters of the environment are learned slowly. These include probabilistic tasks – for example if we observe a binary variable (say, reward vs. no reward), we need several trials to estimate the underlying probability p(reward): to tell the difference between a reward probability of 0.8 and 0.9 would require at least 10 trials for example.

The more the participant's internal model of the environment differs from the true state of the environment, the more useful it is for the experimenter to have a model of what the participant knows/believes about the state of the world rather than assuming the true parameters of the environment are known.

We will now consider a Bayesian computer participant in a one-armed bandit task. This is a task in which learning naturally requires a larger number of trials and hence participants' model of the environment is likely to differ from the true state of the environment. We will see that in this task the Bayesian computer participant can give us rich insights into what participants might think/believe on each trial of the task.

In the one-armed bandit task, participants must choose between two actions, A and B (say, press a button with the left or right hand), only one of which would lead to delivery of a reward. The probability that action A is rewarded is set at some value q; the probability that action B would be rewarded is then (1-q); formally we can say that the probability that action A is rewarded follows a Bernoulli distribution (a single-trial binomial distribution) with probability parameter q. From time to time during the task, the value of q changes to a new value; participants do not know when these changes will occur or how frequently. Hence the participant's task is to infer both the current value of q, and the probability v of a change in q, from the observed data. The details of this model are not central to our point here, which is to illustrate that a Bayesian model can give rich insights into the internal thought processes of the participant. However, for the interested reader we describe the model used to generate the figures in Appendix A.

Figure 4 (left hand panel) illustrates the task data and model fit. Values of q were generated randomly with a jump probability (true value of v) of $1/15$ - the true value of q on each trial is indicated by the white line. The side that was actually rewarded on each trial is indicated by the presence of a dot on the left or right of the plot, respectively. Remember that as p(Left rewarded) = 1-p(Right rewarded), the player knows which side was rewarded on all trials, even when he chose the unrewarded side.

<Please insert Figure 4 about here>

The red line and shaded areas represent the model's maximum likelihood estimate of the state of the environment (the value of q). The shading represents the probability density distribution over q on each trail, according to the model.

Inspecting the model's estimates of the environment across trials, we can see a number of interesting features. Firstly, we notice that the maximum likelihood estimate is close to the true value of q most of the time. However, when there are changes in the underlying environment, the model takes a few trials to 'catch up'. Secondly, we can see that the model's uncertainty about the state of the world generally decreases over time (the shaded area gets narrower over time), but uncertainty increases when there is a change in the environment, or when a change is suspected.

In the right hand panels of Figure 4 (labeled a, b, and c), we take a closer look at the probability density distributions across parameter space for three sets of trials around change points or suspected change points. The data points in question are labeled a, b, and c in the left hand panel. For each data point we show the distribution of probability density across parameter space on that trial and surrounding trials (right hand panel). These plots are analogous to the parameter space plots in the left hand panel of Figure 3, but instead of plotting the distribution of probability density across values of μ and σ^2 we are now plotting probability density across values of q and v.

Just before time point a, the model 'thinks' that q, the probability of the left side being rewarded, is near to 100%, as it has just experienced a long run of left-rewarded trials. At point a, a right-rewarded trial is observed (the actual trial labeled a in the left hand panel is the same one labeled a in the right hand panel). The probability that associated with values of q and v other than those which were favored before point a increases. However, subsequent trials continue to be left-rewarded, and the model reverts to its previous state of believing the probability of left-rewarded trials to be very high.

In contrast, time point b represents a successful update. Prior to b, there was a long run of right rewarded trials, followed by an actual change in q (white line) and a series of left-rewarded trials starts. In this case, the model updates its estimate of q over a series of trials. On trial b itself, the model is clearly entertaining both the

hypothesis that q has changed, and the hypothesis that q remains the same. Note that the 'change' hypothesis is associated with a higher value of v (the peak is further to the right), compared to the 'no change' hypothesis, as we would expect since v is the probability of change, which is inferred based on the number of change points that were observed.

Finally, point c represents a point at which the model is erroneously updated when there was in fact no change in q. Just before point c, the model 'thinks' q is almost 100%, i.e. only left-rewarded trials can occur. It then observes two right-rewarded trials, leading it to think q has changed to favour right-rewarded trials. However, these two trials are followed by more left-rewarded trials, leading the model to revert to its former hypothesis (favouring the left) but with a more moderate probability value, so q is now nearer to 80% than 100% (indeed, the maximum likelihood estimate of q is now nearer to the true value of q, as seen from the white and red lines on the left-hand panel).

We have briefly described some snapshots of 'interesting behavior' of the Bayesian learning algorithm, in order to illustrate how constructing such a model could allow us to 'peek inside' the mind of a model participant to see how its beliefs about the state of the evolve. We have seen, for example, that learning models can capture lags when even an optimal participant could not yet have adjusted to a change in the environment. We have seen that when a model is fit to the actual data observed by a participant, it can indicate when the participant could mis-estimate the parameters of the environment (such as at point c). We have also seen that Bayesian models can give us insights into internal features of learning such as uncertainty, which may themselves predict neural and/or behavioral data. Hopefully this brief illustration will convince the reader that explicitly modeling the contents of the participant's mind, as with a Bayesian learning model, can generate and refine our predictions about what activity we might find in their brain, beyond what could be achieved by simply relating brain activity to stimuli or responses.

Conclusion

In this chapter, we have discussed the use of Bayesian models in cognitive neuroscience. We have illustrated of the main characteristics of Bayesian models, including the representation of beliefs as probability distributions, the use of priors, and sequential updating of information. These models can be highly predictive of the actual behavior displayed by humans and animals during a variety of tasks. We have looked closely at two learning tasks, one in a stable and one in an unstable environment, and charted how the beliefs of a Baysian model change over trials. The parameters of such a model can then be used to interrogate behavior and brain function.

Appendix A: one-armed bandit model

We can write down the *generative* model, by which the rewarded action (A or B) is selected as follows:

p(A rewarded on trial *i*)~ **Bernoulli**(q_i)

$$q_i = \begin{cases} q_{i-1} & if\ J = 0 \\ rand(0,1) & if\ J = 1 \end{cases}$$

… where J is a binary variable which determines whether there was a jump in the value of q between trial i-1 and trial i; J itself is determined by

$J \sim$ **Bernoulli**(v)

… where v is the probability of a jump, e.g. if a jump occurs on average every 15 trials, $v = 1/15$.

Then we can construct a Bayesian computer participant which infers the values of q and v on trial *i* as follows:

$$p(q, v|x_{1:i}) = p(x_i|q_i, v)\, p(q_i, v)$$

where the prior at trial *i*, $p(q_i, v)$, is given by

$$p(q_i, v) = p(q_i|q_{i-1}, v)\, p(q_{i-1}, v|x_{1:i-1})$$

and the transition function $p(q_i|q_{i-1}, v)$ is given by

$$p(q_i|q_{i-1}, v) = (1 - v)q_{i-1} + v\left(\frac{1}{Uniform(0,1)}\right)$$

Exercises

Exercise 1. Look at Figure 5. How do you interpret the shadow on the surface shapes? Most people see the left hand side bumps as convex and the right hand bumps as concaves. Can you explain why that might be, using your Bayesian perspective? Hint: think of the use of priors.

Exercise 2. In figure 4 we saw some interesting behavior by a Bayesian learner. For instance, at point c the model very quickly changed its belief of an environment where left was rewarded into one where right was rewarded. One important goal of model-based cognitive neuroscience is to link this type of changes probability distributions to observed neural phenomena. Can you come up with some phenomena that can be linked with changes in the model's parameters?

<Please insert Figure 5 about here>

Exercise 3. In this final exercise we will ask you to construct a simple Bayesian model. The solutions include example Matlab code, although they are platform independent. Consider the following set of observations of apple positions *x*, which Isaac made in his garden:

i	x_i
1	63
2	121
3	148
4	114

5	131
6	121
7	90
8	108
9	76
10	126

(a) Find the mean, $E(x)$, and variance, $E(x^2)-E(x)^2$, of this set of observations using the formulae

$$E(x) = \frac{1}{n}\sum_i x_i$$

$$E(x^2) = \frac{1}{n}\sum_i x_i^{\,2}$$

(b) If I tell you that these samples were drawn from a normal distribution, $x \sim N(\mu, \sigma^2)$ how could you use Bayes' theorum to find the mean and variance of x? Or more precisely, how could you use Bayes' theorem to estimate the parameters, μ and σ^2, of the normal distribution from which the samples are drawn?

Hint: remember from the text that we can write

$$p(x \sim N(\mu,\sigma^2) | x_1 \ldots x_n) \propto p(x_1 \ldots x_n \mid x \sim N(\mu,\sigma^2))\, p(x \sim N(\mu,\sigma^2))$$

…where the likelihood function, $p(x_i \mid x \sim N(\mu, \sigma))$, is given by the standard probability density function for a normal distribution:

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

…and you can assume:

1. The prior probability $p(x \sim N(\mu,\sigma^2))$ is equal for all possible values of μ and σ^2, and

2. The observations are independent samples such that $p(x_i \cap x_j)=p(x_i)p(x_j)$ for all pairs of samples $\{x_i, x_j\}$.

Now use MATLAB to work out the posterior probability for a range of pairs of parameter values μ and σ^2, and find the pair with the highest joint posterior probability. This gives a maximum likelihood estimate for μ and σ^2.

(c) Can you adapt this model to process each data point sequentially, so that the posterior after observation *i* becomes the prior for observation *i+1* ?

Hint: remember from the text that (assuming the underlying values of μ and σ^2 cannot change between observations), we can write:

$$p(x \sim N(\mu,\sigma^2) | x_1 \ldots x_i) \propto p(x_i \mid x \sim N(\mu,\sigma^2))\ p(x \sim N(\mu,\sigma^2) \mid x_1 \ldots x_{i-1})$$

… where the prior at trial *i*, $p(x \sim N(\mu,\sigma^2) \mid x_1 \ldots x_{i-1})$ is the posterior from trial *i-1*.

(d) If you have done parts 2 and 3 correctly, the final estimates of $\{\mu, \sigma^2\}$ should be the same whether you process the data points sequentially, or all at once. Why is this?

Further reading

1. McGrayne [2] provides an historical overview of the development of Bayes' theorem, its applications, and its gradual acceptance in the scientific community;
2. Daniel Wolpert's TED talk (available at http://www.ted.com/talks/daniel_wolpert_the_real_reason_for_brains.html) provides a nice introduction in to consequences of noise in neural systems and the Bayesian way of dealing with it;
3. O'Reilly [20] discusses Bayesian approaches to dealing with changes in the environment and how different types of uncertainty are incorporated into Bayesian models and dealt with in the brain.

4. Nate Silver's book *The signal and the noise* [21] contains some nice example about how humans make predictions and establish beliefs. Silver advocates a Bayesian approach to dealing with uncertainty. It served him very well in the 2012 USA presidential elections, when he correctly predicted for each of the 50 states whether they would be carried by Obama or Romney.

5. David MacKay's book *Information theory, inference, and learning algorithms* [22] is a much more advanced treatment of many of the principle of Bayesian thinking. It is available for free at http://www.inference.phy.cam.ac.uk/itprnn/book.html.

References

1. Bayes T (1763) An essay towards solving a problem in the doctrine of chances. *Phil Trans* 53:370-418.
2. McGrayne SB (2011) *The theory that would not die: How Bayes' rule cracked the enigma code, hunted down Russian submarines, and emerged triumphant from two centuries of controversy.* New Haven: Yale University Press.
3. Cox RT (1946) Probability, frequency and reasonable expectation. *Am J Phys* 14:1-13.
4. O'Reilly JX, Jbabdi S, & Behrens TE (2012) How can a Bayesian approach inform neuroscience? *Eur J Neurosci* 35:1169-1179.
5. Chater N & Oaksford M eds (2008) *The probabilistic mind: Prospects for Bayesian cognitive science* (Oxford University Press, Oxford).
6. Körding KP & Wolpert DM (2006) Bayesian decision theory in sensorimotor control. *Trends Cogn Sci* 10:319-326.
7. Gregory R (1966) *Eye and brain.* Princeton: Princeton University Press.
8. Segall MH, Campbell DT, & Herskovits MJ (1963) Cultural differences in the perception of geometric illusions. *Science* 139:769-771.
9. Jacobs RA (1999) Optimal integration of texture and motion cues to depth. *Vis Res* 39:3621-3629.
10. Ernst MO & Banks MS (2002) Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* 415:429-433.
11. Behrens TE, Woolrich MW, Walton ME, & Rushworth MF (2007) Learning the value of information in an uncertain world. *Nat Neurosci* 10:1214-1221.
12. Robbins H (1952) Some aspects of the sequential design of experiments. *Bull Amer Math Soc* 58:527-535.
13. Real LA (1991) Animal choice behavior and the evolution of cognitive architecture. *Science* 253:980-986.
14. Posner MI, Snyder CRR, & Davidson BJ (1980) Attention and the detection of signals. *J Exp Psychol Gen* 109:160-174.
15. Knight FH (1921) *Risk, uncertainty and profit.* Boston: Hart, Schaffner and Marx.

16. Payzan-LeNestour E & Bossaerts P (2011) Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings. *PLoS Comp Biol* 7:e1001048.
17. Courville AC, Daw ND, & Touretzky DS (2006) Bayesian theories of conditioning in a changing world. *Trends Cogn Sci* 10:294-300.
18. Dayan P, Kakade S, & Montague PR (2000) Learning and selective attention. *Nat Neurosci* 3:1218-1223.
19. O'Reilly JX & Mars RB (2011) Computational neuroimaging: Localising Greek letters? *Trends Cogn Sci* 15:450.
20. O'Reilly JX (in press) Uncertainty, learning and prediction in a changing world. *Front Neurosci*.
21. Silver N (2012) *The signal and the noise: Why most predictions fail but some don't.* New York: Penguin.
22. MacKay DJC (2003) *Information theory, inference, and learning algorithms.* Cambridge: Cambridge University Press.

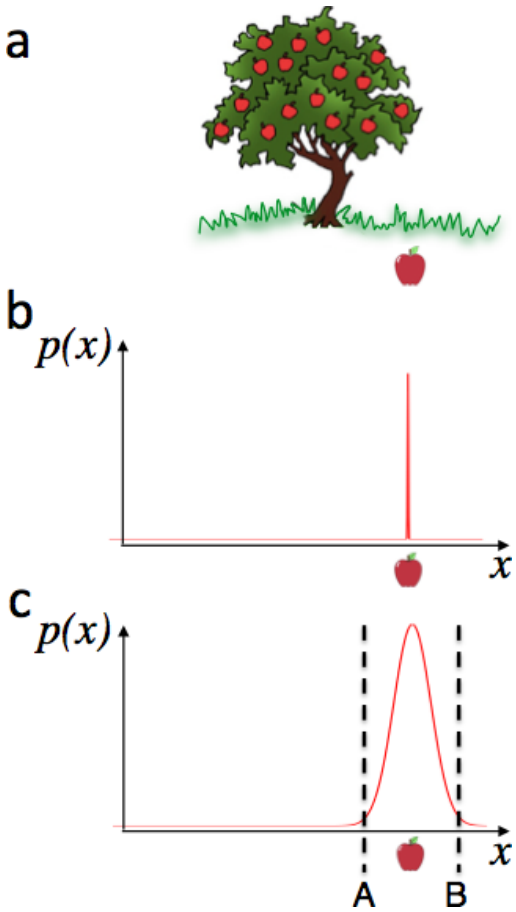

Figure 1.

a. Apple under tree

b. A single value representation of the most likely locations for apples to fall.

c. Probabilistic representation of apple falling positions.

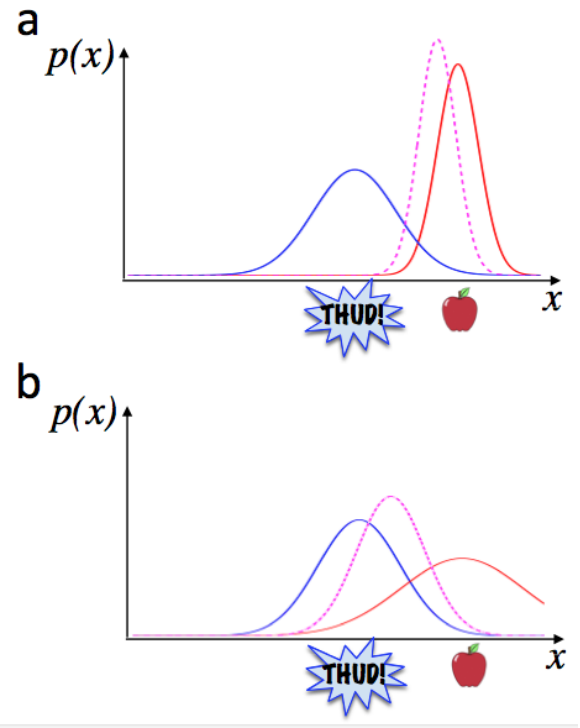

Figure 2.

Multisensory integration

a. The probability density function (pdf) of hearing (blue line) is much wider than that of vision (red line), indicating that hearing is associated with much more uncertainty. As a result, the combined probability density function (dotted line) is closer to the vision pdf.

b. When vision becomes more uncertain, the resulting combined pdf (dotted line) is much closer to that of hearing.

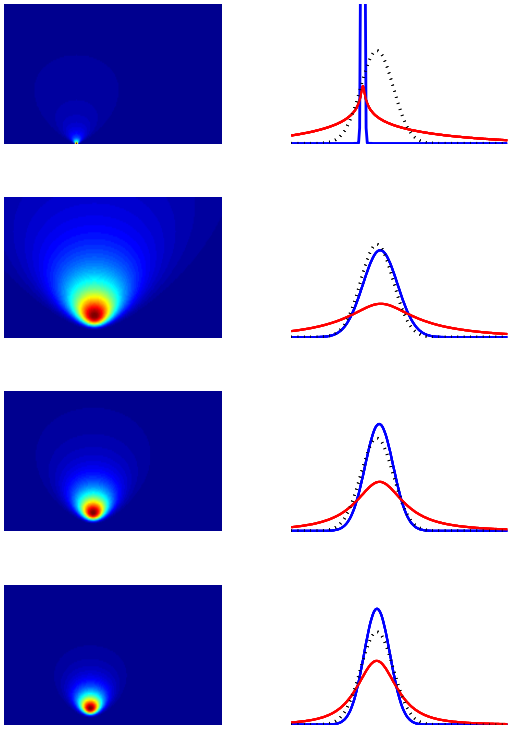

Figure 3.

Four trials of learning the mean and variance of a Gaussian.

The left hand panels indicate the likelihood of each part of state space. The right hand panels indicate the true distribution (dotted line), the maximum likelihood distribution of apple locations (black lines), and the weighted combination of all possible Gaussian distributions over locations (grey lines).

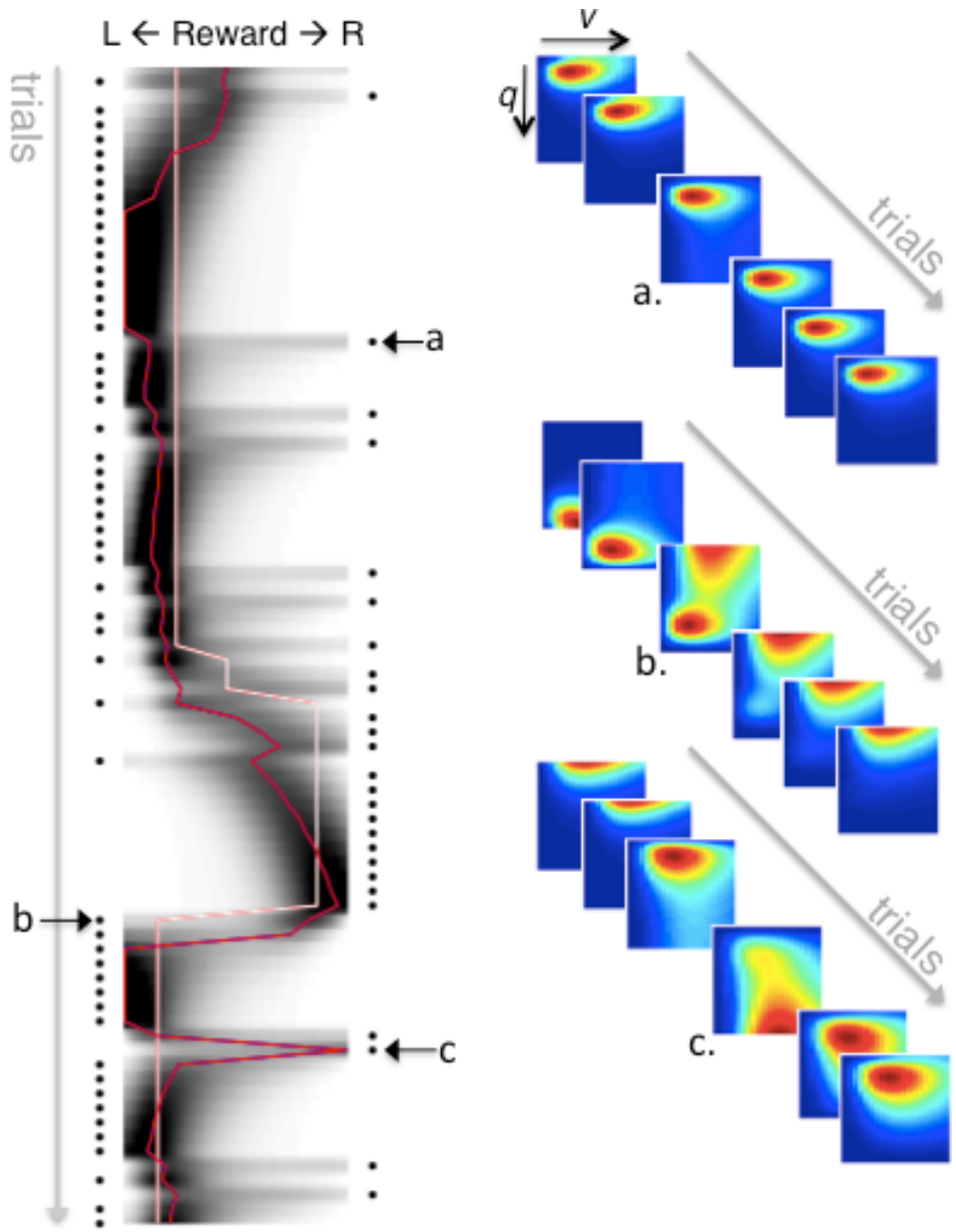

Figure 4.

Learning in an unstable environment.

The left hand panel shows the rewarded sides on each trial (black dots); the true probability of reward, i.e., the true value of *q* (white line); the model's estimate of *q* (red line); and the uncertainty of the model's estimation (shade). The left hand panels show the model's beliefs of each possibility in state space. *(a)* No update - A point at which a single right side reward trial is observed. The probability in the right hand region representing increases, but no further right reward trials occur, and the model goes back to expecting LH to be rewarded. *(b)* Successful update. *(c)* False/temporary update.

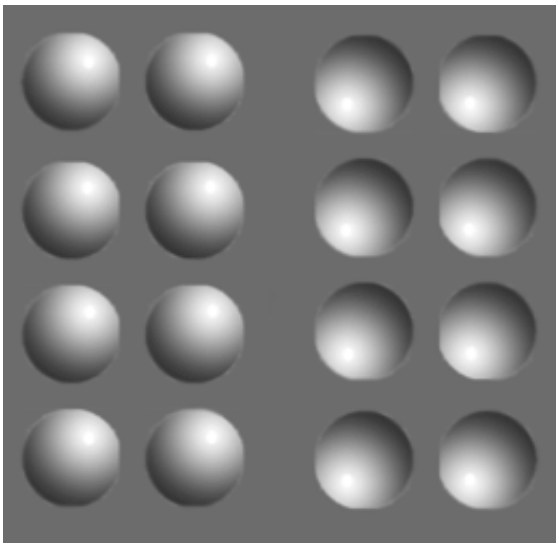

Figure 5.

Convex or concave?