CellPress

Neuron
Primer

Bayesian Decision Models: A Primer

Wei Ji Ma[1,*]
[1]Center for Neural Science and Department of Psychology, New York University, New York, NY, USA
*Correspondence: weijima@nyu.edu
https://doi.org/10.1016/j.neuron.2019.09.037

To understand decision-making behavior in simple, controlled environments, Bayesian models are often useful. First, optimal behavior is always Bayesian. Second, even when behavior deviates from optimality, the Bayesian approach offers candidate models to account for suboptimalities. Third, a realist interpretation of Bayesian models opens the door to studying the neural representation of uncertainty. In this tutorial, we review the principles of Bayesian models of decision making and then focus on five case studies with exercises. We conclude with reflections and future directions.

1. Introduction

1.1. What Are Bayesian Decision Models?

Good computational modeling of decision making goes beyond a mere description of the data (curve fitting). Good models help to break down perceptual, cognitive, or motor processes into interpretable and generalizable stages. This, in turn, may allow for conceptual connections across experiments or domains, for a characterization of individual differences, or for establishing correlations between model variables and aspects of neural activity.

Across domains of application, Bayesian models of decision making are based on the same small set of principles, thereby promising high interpretability and generalizability. Bayesian models aspire to account for an organism's decision process when the task-relevant states of the world are not exactly known to the organism. A state of the world can be a physical variable, such as the reflectance of a surface or the location where a ball will land, or a more abstract one, such as whether two parts belong to the same object or the intent of another person. Lack of exact knowledge can arise from noise (Faisal et al., 2008) or from missing information, such as in the case of occlusion (Kersten et al., 2004).

Bayesian decision models have two key components (Figure 1). The first is Bayes' rule, which formalizes how the decision maker assigns probabilities (degrees of belief) to hypothesized states of the world given a particular set of observations. The second is a cost function, which is the quantity that the decision maker would like to minimize; an example would be the proportion of errors in a task. The cost function dictates how the decision maker should transform beliefs about states of the world into a decision. Combining the components, we end up with a mapping from observations to decision. While the form of that mapping depends on the task, the Bayesian recipe for deriving it is general.

1.2. Why Bayesian Decision Models?

The Bayesian modeling framework for decision making holds appeal for various reasons. The first reason has an evolutionary or ecological flavor: Bayesian inference optimizes behavioral performance, and one might postulate that the mind applies a near-optimal algorithm in decision tasks that are common or important in the natural world (or daily life). This argument is more plausible for perceptual than for cognitive decision making. Second, Bayesian models are general because its two key components are; the recipe for constructing a Bayesian model applies across a wide range of tasks. Third, in Bayesian models, the decision model is largely dictated by the generative model, which, in turn, is often largely dictated by the statistics of the experiment. As a result, many Bayesian models have few free parameters. Fourth, Bayesian models have a good empirical track record of accounting for behavior, both in humans and in other animals. Fifth, sensible models can easily be constructed by modifying the assumptions of an optimal Bayesian model. Thus, the Bayesian model is a good starting point for model generation.

1.3. What Math Does One Need?

Bayesian modeling may seem intimidating to beginners, but the math involved rarely goes beyond standard calculus. Because of the recipe-based methodology, many learners already feel in control after several tens of hours of practice. Importantly, when building Bayesian models, it is easy to supplement math and intuitions with simple simulations.

1.4. Areas of Application

Bayesian modeling is most straightforward if the task-relevant world states are categorical (e.g., binary) or low dimensional, if their distributions are simple, if the task allows for parametric variation of world state variables, and if the decision maker has an unambiguously defined objective. This makes Bayesian modeling, in particular, suitable for simple perceptual tasks. However, Bayesian decision models appear also in studies of eye movements in natural scenes (Itti and Baldi, 2009), reaching movements (Körding and Wolpert, 2004), physical scene understanding (Battaglia et al., 2013), speech understanding (Goodman and Frank, 2016), inductive reasoning (Tenenbaum et al., 2006), and economic decisions (Cogley and Sargent, 2008).

1.5. Disclaimer

This primer is about Bayesian decision models in psychology and neuroscience, not about Bayesian data analysis. We will also not discuss how to fit Bayesian decision models and compare them to other decision models because that involves methods that are not specific to Bayesian modeling and are well described elsewhere.

2. Recipe for Bayesian Modeling

The recipe for Bayesian modeling consists of the following steps: specifying the "generative model" (step 1), calculating the

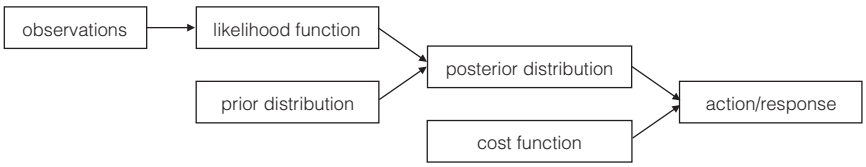


Figure 1. Schematic of Bayesian Decision Making

"posterior distribution" (inference; step 2a), turning the posterior distribution into an "action or response" (step 2b), and calculating the "action/response distribution" for comparison with experimental data (step 3).

2.1. Step 1: Generative Model

Consider a decision maker who has to take an action that requires inferring a state of the world s from an observation x. The variables s and x can be discrete or continuous and one-dimensional or multidimensional. The observation x can be a physical stimulus generated from an underlying unknown s (for example, when s is a category of stimuli), an abstract "measurement" (s plus noise), or a pattern of neural activity.

The frequencies of occurrence of each value of s in the environment are captured by a probability distribution $p(s)$. The distribution of the observation is specified conditioned on s and denoted by $p(x \mid s)$. Often, $p(x \mid s)$ is not directly known but has to be derived from other conditional distributions. Together, the distributions $p(s)$ and $p(x \mid s)$ define a "generative model," a statistical description of how the observations come about.

2.2. Step 2a: Inference

The key assumption of Bayesian models of behavior is that the decision maker has learned the distributions in the generative model and puts this knowledge to full use when inferring states of the world. We now turn to this inference process.

On a given trial, the decision maker makes a specific observation x_{trial}. They then calculate the probabilities of possible world states s given that observation. They do this using Bayes' rule,

$$p(s \mid x_{\text{trial}}) = \frac{p(x_{\text{trial}} \mid s)p(s)}{p(x_{\text{trial}})}. \qquad \text{(Equation 1)}$$

The numerator of the right-hand side involves two probabilities that we recognize from the generative model. Indeed, because we defined them in the generative model, they can be calculated here. However, their interpretation is different from the generative model. To understand this, we first need to realize that, in Equation 1, the world state variable s should be considered a hypothesis entertained by the decision maker. Each probability involving s should be interpreted as a degree of belief in a value of s. As such, these probabilities exist only in the head of the decision maker and are not directly observable.

Specifically, in the context of Equation 1, $p(s)$ is called the "prior distribution." At first glance, it might confusing to give $p(s)$ a new name; was it not already the distribution of the state of the world? The reason for the new name is that the prior distribution reflects to what extent the decision maker *expects* different values of s; in other words, it formalizes a belief. If the decision maker's beliefs are wrong, the distribution $p(s)$ in Equation 1 will be different from the $p(s)$ in the generative model; we will discuss this in Section 4.5.

The factor $p(x_{\text{trial}} \mid s)$ in Equation 1 is the likelihood function of s. The likelihood of a hypothesized world state s given an observation is the probability of those observations if that hypothesis were true. It is important that the likelihood function is a function of the hypothesized world state s, not of x_{trial} (which is a given value). To make this dependence explicit, it could be helpful to use the notation $L(s) = p(x_{\text{trial}} \mid s)$. The likelihood of s is numerically the same as $p(x_{\text{trial}} \mid s)$ in the generative model, but what is the argument and what is given is switched. (Side note: Bayesian experts would not use the phrase "likelihood of the observation.")

The left-hand side distribution, $p(s \mid x_{\text{trial}})$, is called the posterior distribution of s. It captures the answer to the question to what extent each possible world state value s is supported by the observed measurement x_{trial} and prior beliefs. Finally, the probability in the denominator of Equation 1, $p(x_{\text{trial}})$, does not depend on s and is therefore a numerical constant; it acts as the normalization factor of the numerator.

2.3. Step 2b: Taking an Action (Making a Response)

Bayesian decision making does not end with the computation of a posterior distribution. The decision maker has to take an action, which we will denote by a. An action could be a natural movement or a response in an experiment. In perceptual tasks in the laboratory, the response is typically an estimate $\hat{s}$ of the state of the world s, and such an estimate is Bayesian if it is derived from the posterior. In perceptual tasks, one could even go as far as postulating that $\hat{s}$ represents the contents of perception (i.e., a percept).

In general, the appropriate action a can be chosen in a principled manner by defining a "cost function" $C(s,a)$ (also called a "loss function" or an "objective function"), which the decision maker strives to minimize. This function depends on the world state s and the action a. On a trial when the observation is x_{trial}, the "expected cost" is the expected value of $C(s,a)$ with respect to the posterior distribution calculated in Equation 1:

$$\mathbb{E}C(a) = \sum_s p(s \mid x_{\text{trial}})C(s,a), \qquad \text{(Equation 2)}$$

where the sum applies to discrete s; for continuous s, the sum is replaced by an integral. The Bayesian decision maker chooses the action that minimizes $\mathbb{E}C(a)$. In this sense, the action is *optimal* and the Bayesian approach is *normative*. The process of choosing an action given a posterior is, in its basic form, deterministic.

When a is an estimate of s, we can be a bit more specific. First, we consider the case that s is discrete, and the objective is to estimate correctly (i.e., $C(s,a) = -1$ when $s = a$, 0 otherwise). Then, the optimal action is to report the mode of the posterior: the state for which $p(s \mid x_{\text{trial}})$ is highest. Next, we consider the case s is real valued, and the objective is to minimize expected squared error (i.e., $C(s,a) = (s-a)^2$). Then, the optimal readout is the mean of the posterior distribution.

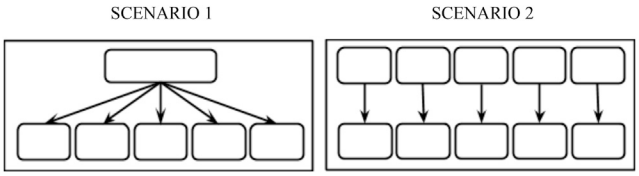


Figure 2. Generative Model Diagram for the Two Scenarios in Case 1

2.4. Step 3: Response (Action) Distribution

The generative model, the computation of the posterior (inference), and a mapping from posterior to action together complete a Bayesian model. To test the model against experimental data, the modeler needs to derive or simulate the distribution of actions a given a state of the world s—that is, $p(a \mid s)$. When the action is an estimate $\hat{s}$, the difference $\hat{s} - s$ is the estimation error, and the distribution $p(\hat{s} \mid s)$ will characterize estimation errors. Finally, the parameters of the model need to be fitted to the data. Examples of parameters in Bayesian models are sensory noise level (Section 3.5), lapse rate (Section 4.4), and wrong belief parameters (Section 4.5). For fitting, maximum-likelihood estimation is the most standard, where now "likelihood" refers to the likelihood of the parameters given the experimental data (Myung, 2003). To implement the maximization, one can use any number of standard algorithms; make sure to initialize with multiple starting points to reduce the chance of getting stuck in a local optimum (Martí et al., 2016).

3. Case Studies

We will now go through five case studies that illustrate different aspects of Bayesian decision models. We encourage the reader to try to do the exercises; solutions are included in Methods S1.

3.1. Case 1: Unequal Likelihoods and Gestalt Laws

You observe the five dots below all moving downward, as indicated by the arrows.

According to Gestalt psychology (Wertheimer, 1938), the mind has a tendency to group the dots together because of their common motion and perceive them as a single object. This is captured by the "Gestalt law of common fate." Gestalt laws, however, are merely narrative summaries of phenomenology. A Bayesian model has the potential to provide a true explanation of the percept and, in some cases, make quantitative predictions (Wagemans et al., 2012). In this case, the Bayesian decision model takes the form of an "observer model" or "perception model."

Step 1: Generative Model

We first formulate our generative model. The retinal image of each dot serves as a sensory observation. We will denote these five retinal images by I_1, I_2, I_3, I_4, and I_5, each specifying the direction of movement of the corresponding dot's image on the retina (up or down). For didactic purposes, let's say that there exist only two scenarios in the world.

- Scenario 1: all dots are part of the same object, and they therefore always move together. They move together either up or down, each with probability 0.5.
- Scenario 2: each dot is an object by itself. Each dot independently moves either up or down, each with probability 0.5.

(Dots are only allowed to move up and down, and speed and position do not play a role in this problem.) The world state s from Section 2 is now a binary scenario.

a. The generative model diagram in Figure 2 shows each scenario in a big box. Inside each box, the bubbles contain the variables and the arrows represent dependencies between variables. In other words, an arrow can be understood to represent the influence of one variable on another; it can be read as "produces" or "generates" or "gives rise to." The sensory observations should always be at the bottom of the diagram. Put the following variable names in the correct boxes: retinal images I_1, I_2, I_3, I_4, and I_5 and motion directions s (a single motion direction), s_1, s_2, s_3, s_4, and s_5. The same variable might appear more than once.

Step 2: Inference

In inference, the two scenarios become hypothesized scenarios. Inference involves likelihoods and priors. The likelihood of a scenario is the probability of the sensory observations under the scenario.

b. What is the likelihood of scenario 1?
c. What is the likelihood of scenario 2?
d. Do the likelihoods of the scenarios sum to 1? Explain why or why not.
e. What is wrong with the phrase "the likelihood of the observations"?

Let's say scenario 1 occurs twice as often in the world as scenario 2. The observer can use these frequencies of occurrence as prior probabilities, reflecting expectations in the absence of specific sensory observations.

f. What are the prior probabilities of scenarios 1 and 2?
g. What is the product of the likelihood and the prior probability for scenario 1?
h. What is this product for scenario 2?
i. Do these products of the scenarios sum to 1?
j. Posterior probabilities have to sum to 1. To achieve that, divide each of the products above by their sum. Calculate the posterior probabilities of scenarios 1 and 2. You have just applied Bayes' rule.

The default Bayesian perception model for discrete hypotheses holds that the percept is the scenario with the highest posterior probability (*maximum-a-posteriori* or MAP estimation).

k. Would that be consistent with the law of common fate? Explain.

l. How does this Bayesian observer model complement—or go beyond—the traditional Gestalt account of this phenomenon?

In this case, like often, the action is in the likelihood and the prior is relatively unimportant.

3.2 Case 2: Competing Likelihoods and Priors in Motion Sickness

Michel Treisman has tried to explain motion sickness in the context of evolution (Treisman, 1977). During the millions of years over which the human brain evolved, accidentally eating toxic food was a real possibility, and that could cause hallucinations. Perhaps, our modern brain still uses prior probabilities passed on from those days; those would not be based on our personal experience, but on our ancestors'! This is a fascinating, though only weakly tested, theory. Here, we don't delve into the merits of the theory but try to cast it in Bayesian form.

Suppose you are in the windowless room on a ship at sea. Your brain has two sets of sensory observations: visual observations and vestibular observations. Let's say that the brain considers three scenarios for what caused these observations:

- Scenario 1: the room is not moving and your motion in the room causes both sets of observations.
- Scenario 2: your motion in the room causes your visual observations, whereas your motion in the room and the room's motion in the world together cause the vestibular observations.
- Scenario 3: you are hallucinating; your motion in the room and ingested toxins together cause both sets of observations.

Step 1: Generative Model

a. Draw a diagram of the generative model. It should contain one box for each scenario, and all of the italicized variables in the previous paragraph. Some variables might appear more than once.

Step 2: Inference

No numbers are needed except in part (e).

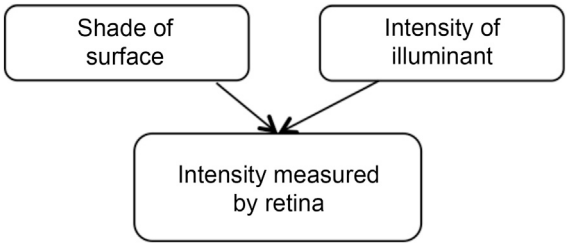
b. In prehistory, people would, of course, move around in the world, but surroundings would almost never move. Once in a while, a person might accidentally ingest toxins. Assuming that your innate prior probabilities are based on these prehistoric frequencies of events, draw a bar diagram to represent your prior probabilities of the three scenarios above.
c. In the windowless room on the ship, there is a big discrepancy between your visual and vestibular observations. Draw a bar diagram that illustrates the *likelihoods* of the three scenarios in that situation (i.e., how probable these particular sensory observations are under each scenario).
d. Draw a bar diagram that illustrates the *posterior probabilities* of the three scenarios.
e. Use numbers to illustrate the calculations in (b)–(d).
f. Using the posterior probabilities, explain why you might vomit in this situation.

3.3. Case 3: Ambiguity Due to a Nuisance Parameter in Color Vision

We switch domains once again and apply a Bayesian approach to the central problem of color vision (Brainard and Freeman, 1997), simplified to a problem for grayscale surfaces. We see a surface when there is a light source. The surface absorbs some proportion of the incident photons and reflects the rest. Some of the reflected photons reach our retina.

Step 1: Generative Model

The diagram of the generative model is:



The "shade" of a surface is the grayscale in which a surface has been painted. Technically, shade is "reflectance," the proportion of incident light that is reflected. Black paper might have a reflectance of 0.10, while white paper might have a reflectance of 0.90. The "intensity of a light source" (illuminant) is the amount of light it emits. Surface shade and light intensity are the world state variables relevant to this problem.

The sensory observation is the amount of light measured by the retina, which we will also refer to as retinal intensity. The retinal intensity can be calculated as follows:

$$\text{Retinal intensity} = \text{surface shade} \times \text{light intensity} \quad \text{(Equation 3)}$$

In other words, if you make a surface twice as reflectant, it has the same effect on your retina as doubling the intensity of the light source.

Step 2: Inference

Let's take each of these numbers to be between 0 (representing black) and 1 (representing white). For example, if the surface shade is 0.5 (mid-level gray) and the light intensity is 0.2 (very dim light), then the retinal intensity is $0.5 \times 0.2 = 0.1$.

a. Suppose your retinal intensity is 0.2. Suppose further that you hypothesize the light intensity to be 1 (very bright light). Under that hypothesis, calculate what the surface shade must have been.
b. Suppose your retinal intensity is the same 0.2. Suppose further that you hypothesize the light intensity to be 0.4. Under that hypothesis, calculate what the surface shade must have been.
c. Explain why the retinal intensity provides *ambiguous* information about surface shade.
d. Suppose your retinal intensity is again 0.2. By going through a few more examples like the ones in (a) and (b), draw in the two-variable likelihood diagram in Figure 3 all combinations of hypothesized surface shade and hypothesized light intensity that could have produced your retinal intensity of 0.2. Think of this plot as a 3D plot (surface plot)!

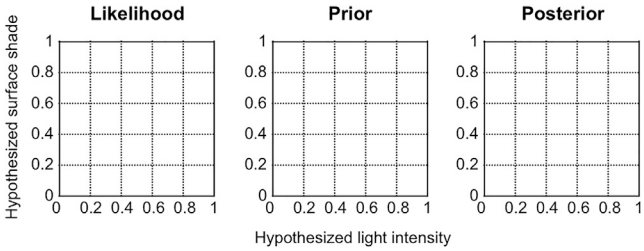


Figure 3. Two-Variable Diagrams for Case 3

e. Explain the statement: "The curve that we just drew represents the combinations of surface shade and light intensity that have a high likelihood."
f. Suppose you have a strong prior that light intensity was between 0.2 and 0.4 and definitely nothing else. In the two-variable prior diagram in Figure 3 (center), shade the area corresponding to this prior.
g. In the two-variable posterior diagram in Figure 3 (right), indicate where the posterior probability is high.
h. What would you perceive according to the Bayesian theory?

3.4. Case 4: Inference under Measurement Noise in Sound Localization

The previous cases featured categorically distinct scenarios. We now consider a continuous estimation task—for example, locating a sound on a line. This will allow us to introduce the concept of noise in the internal measurement of the stimulus. This case would be uninteresting without such noise.

Step 1: Generative Model

The stimulus is the location of the sound. The sensory observations generated by the sound location consist of a complex pattern of auditory neural activity, but for the purpose of our model, and reflecting common practice, we reduce the sensory observations to a single scalar, namely a noisy internal measurement x. The measurement lives in the same space as the stimulus itself—in this case, the real line. For example, if the true location s of the sound is 3° to the right of straight ahead, then its measurement x could be 2.7° or 3.1°.

Thus, the problem contains two variables: the stimulus s and the observer's measurement x. Each node in the graph is associated with a probability distribution: the stimulus node with a stimulus distribution $p(s)$ and the measurement node with a measurement distribution $p(x\,|\,s)$ that depends on the value of the stimulus. In our example, say that the experimenter has programmed $p(s)$ to be Gaussian with a mean μ and variance σ_s^2.

$$p(s) = \frac{1}{\sqrt{2\pi\sigma_s^2}} e^{-\frac{(s-\mu)^2}{2\sigma_s^2}}. \qquad \text{(Equation 4)}$$

(See Figure 4A.) The "measurement distribution" is the distribution of the measurement x for a given stimulus value s. We make the common assumption that the measurement distribution is Gaussian:

$$p(x\,|\,s) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-s)^2}{2\sigma^2}}, \qquad \text{(Equation 5)}$$

where σ is the standard deviation of the measurement noise, also called "measurement noise level" or "sensory noise level." This Gaussian distribution is shown in Figure 4B. The higher σ, the noisier the measurement and the wider its distribution. The Gaussian assumption can be justified using the Central Limit Theorem.

Step 2a: Inference

On a given trial, the observer makes a measurement x_{trial}. The inference problem is: what stimulus estimate should the observer make?

We introduced the stimulus distribution $p(s)$, which reflects how often each stimulus value tends to occur in the experiment. Suppose that the observer has learned this distribution through training. Then, the observer will already have an expectation about the stimulus before it even appears. This expectation constitutes prior knowledge, and, therefore, in the inference process, $p(s)$ is referred to as the "prior distribution" (Figure 5A). Unlike the stimulus distribution in the generative model, the prior distribution reflects the observer's beliefs. The likelihood function represents the observer's belief about the stimulus based on the measurement only—absent any prior knowledge. Formally, the likelihood is the probability of the observed measurement under a hypothesized stimulus:

$$L(s) = p(x_{\text{trial}}\,|\,s). \qquad \text{(Equation 6)}$$

As stated in Section 2.2, the likelihood function is a function of s, not of x. The x variable is now fixed to the observed value x_{trial}. Under our assumption for the measurement distribution $p(x\,|\,s)$, the likelihood function over the stimulus is

$$L(s) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(s-x_{\text{trial}})^2}{2\sigma^2}}. \qquad \text{(Equation 7)}$$

(Although this particular likelihood is normalized over s, that is not generally true. This is why the likelihood function is called a *function* and not a distribution.) The width of the likelihood function is interpreted as the observer's level of *uncertainty* based on the measurements alone.

The posterior distribution is $p(s\,|\,x_{\text{trial}})$, the probability density function over the stimulus variable s given the measurement x_{trial}. We rewrite Bayes' rule, Equation 1, as

$$p(s|x_{\text{trial}}) \propto p(x_{\text{trial}}\,|\,s)p(s) = L(s)p(s). \qquad \text{(Equation 8)}$$

a. Why can we get away with the proportionality sign?

Equation 8 assigns a probability to each possible *hypothesized value* of the unknown stimulus s. We will now compute the posterior distributions under the assumptions we made in step 1. Upon substituting the expressions for $L(s)$ and $p(s)$ into

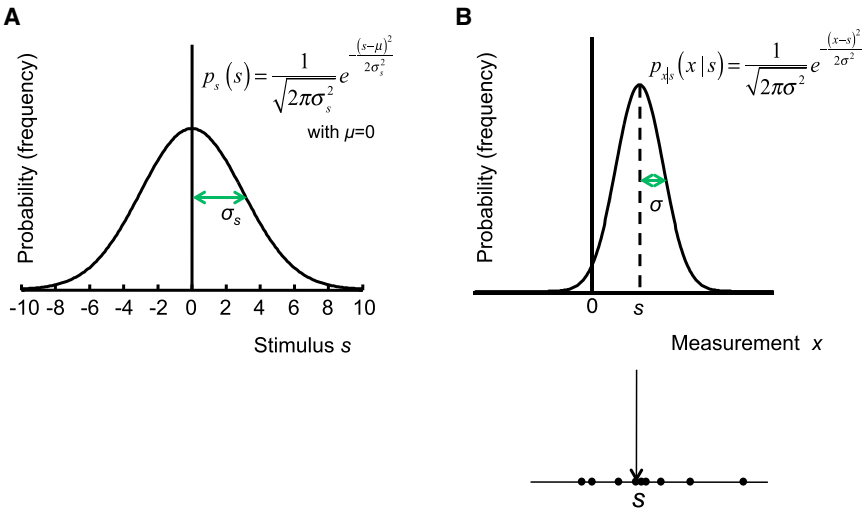


Figure 4. The Probability Distributions that Belong to the Two Variables in the Generative Model

(A) A Gaussian distribution over the stimulus, $p(s)$, reflecting the frequency of occurrence of each stimulus value in the world.

(B) Suppose we now fix a particular value of s (the dotted line). Then, the measurements x will follow a Gaussian distribution around that s. The diagram at the bottom shows a few samples of x, which are scattered around the true sound location s, indicated by the arrow.

Equation 8, we see that in order to compute the posterior, we need to compute the product of two Gaussian functions. An example is shown in Figure 6.

b. Create a figure similar to Figure 6 through numerical computation of the posterior. Numerically normalize prior, likelihood, and posterior.

Beyond plotting the posterior, our assumptions in this case actually allow us to characterize the posterior mathematically.

c. Show that the posterior is a new Gaussian distribution

$$p(s \mid x_{\text{trial}}) = \frac{1}{\sqrt{2\pi\sigma^2_{\text{posterior}}}} e^{\frac{(s-\mu_{\text{posterior}})^2}{2\sigma^2_{\text{posterior}}}}, \qquad \text{(Equation 9)}$$

with mean

$$\mu_{\text{posterior}} = \frac{\frac{x_{\text{trial}}}{\sigma^2} + \frac{\mu}{\sigma_s^2}}{\frac{1}{\sigma^2} + \frac{1}{\sigma_s^2}} \qquad \text{(Equation 10)}$$

and variance

$$\sigma^2_{\text{posterior}} = \frac{1}{\frac{1}{\sigma^2} + \frac{1}{\sigma_s^2}}. \qquad \text{(Equation 11)}$$

You may use the following auxiliary calculation:

$$-\frac{(s-\mu)^2}{2\sigma_s^2} - \frac{(s-x_{\text{trial}})^2}{2\sigma^2} = -\frac{1}{2}\left(\frac{1}{\sigma_s^2} + \frac{1}{\sigma^2}\right)\left(s - \frac{\frac{\mu}{\sigma_s^2} + \frac{x_{\text{trial}}}{\sigma^2}}{\frac{1}{\sigma_s^2} + \frac{1}{\sigma^2}}\right)^2 + \text{junk}$$

where "junk" refers to terms that don't depend on s (see part a to understand why we can ignore these when calculating the new distribution).

The mean of the posterior, Equation 10, is of the form $ax_{\text{trial}} + b\mu$; in other words, it is a linear combination of x_{trial} and the mean of the prior, μ. The coefficients a and b in this linear combination are $((1/\sigma^2)/(1/\sigma^2 + 1/\sigma_s^2))$ and $((1/\sigma_s^2)/(1/\sigma^2 + 1/\sigma_s^2))$, respectively. These sum to 1, and, therefore, the linear combination is a "weighted average," where the coefficients act as weights. This weighted average, $\mu_{\text{posterior}}$, will always lie somewhere in between x_{trial} and μ.

d. In the special case that $\sigma = \sigma_s$, compute the mean of the posterior.

The intuition behind the weighted average is that the prior "pulls the posterior away" from the measurement x_{trial} and toward its own mean μ, but its ability to pull depends on how narrow it is compared to the likelihood function. If the likelihood function is narrow, which happens when the noise level σ is low, the posterior won't budge much: it will be centered close to the mean of the likelihood function. This intuition is still valid if the likelihood function and the prior are not Gaussian but are roughly bell shaped.

The variance of the posterior is given by Equation 11. It is interpreted as the overall level of uncertainty the observer has about the stimulus after combining the measurement with the prior. It is different from both the variance of the likelihood function and the variance of the prior distribution.

e. Show that the variance of the posterior can also be written as $\sigma^2_{\text{posterior}} = \sigma^2\sigma_s^2/(\sigma^2 + \sigma_s^2)$.

f. Show that the variance of the posterior is smaller than both the variance of the likelihood function and the variance of the prior distribution. This shows that combining a measurement with prior knowledge makes an observer less uncertain about the stimulus.

g. What is the variance of the posterior in the special case that $\sigma = \sigma_s$?

Step 2b: The Stimulus Estimate (response)

We now estimate s on the trial under consideration. We denote the estimate by $\hat{s}$. As mentioned in Section 2.3, for a real-valued variable and a squared error loss function, the observer should use the mean of the posterior as the estimate. Thus,

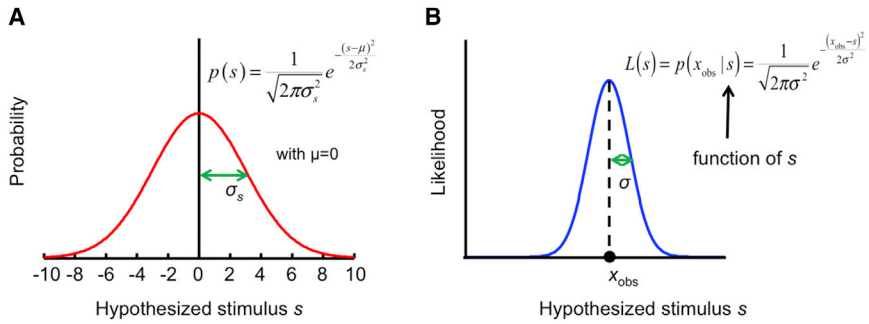


Figure 5. Prior and Likelihood under Sensory Noise

Consider a single trial on which the observed measurement is x_{trial}. The observer is trying to infer which stimulus s produced this measurement. The two functions that play a role in the observer's inference process (on a single trial) are the prior and the likelihood. The argument of both the prior and the likelihood function is s, the hypothesized stimulus.

(A) Prior distribution with $\mu = 0$. This distribution reflects the observer's beliefs about different possible values the stimulus can take.

(B) The likelihood function over the stimulus based on the measurement x_{trial}. Under our assumptions, the likelihood function is a Gaussian centered at x_{trial}.

$$\hat{s} = \mu_{posterior} = \frac{\frac{x_{trial}}{\sigma^2} + \frac{\mu}{\sigma_s^2}}{\frac{1}{\sigma^2} + \frac{1}{\sigma_s^2}} \quad \text{(Equation 12)}$$

This would be a Bayesian observer's response in this localization task.

Step 3: Response Distribution

We now like to use this model to predict subjects' behavior in this experiment. To do so, we'd like to compare our predicted responses, $\hat{s}$, to the subject's actual responses. Looking at Equation 12 for $\hat{s}$, we note that, to compute a predicted response on a given trial, we need to know x_{trial}. But this is something we don't know! x_{trial} is the noisy measurement made by the observer's sensory system, an internal variable to which an experimenter has no access.

A common mistake in Bayesian modeling is to discuss the likelihood function (or the posterior distribution) as if it were a single, specific function in a given experimental condition. In the presence of noise in the observation/measurement, this is incorrect. Both the likelihood and the posterior depend on the measurement x_{trial}, which itself is randomly generated on each trial, and, therefore, the likelihood and posterior will "wiggle around" from trial to trial (Figure 7). This variability propagates to the estimate: the estimate $\hat{s}$ also depends on the noisy measurement x_{trial} via Equation 12. Since x_{trial} varies from trial to trial, so does the estimate: the stochasticity in the estimate is *inherited from* the stochasticity in the measurement x_{trial}. Hence, in response to repeated presentations of the same stimulus, the estimate will be a random variable with a probability distribution, which we will denote by $p(\hat{s}|s)$.

So rather than comparing our model's predicted responses to subjects' actual responses on individual trials, we'll instead use our model to predict the distribution over subjects' responses for a given value of the stimulus. The predicted distribution is precisely $p(\hat{s}|s)$. To compare our Bayesian model with an observer's behavior, we thus need to calculate this distribution.

h. From step 1, we know that when the true stimulus is s, x_{trial} follows a Gaussian distribution with mean s and variance σ^2. Show that when the true stimulus is s, the estimate's distribution $p(\hat{s}|s)$ is a Gaussian distribution with mean $\frac{s}{\sigma^2} + \frac{\mu}{\sigma_s^2} \Big/ \left(\frac{1}{\sigma^2} + \frac{1}{\sigma_s^2}\right)$ and variance $\frac{1}{\sigma^2} \Big/ \left(\frac{1}{\sigma^2} + \frac{1}{\sigma_s^2}\right)^2$.

We see that the variance of the estimate can be different from the variance of the posterior. Intuitively, the response distribution (the distribution of the observer's posterior mean estimate) for a given true stimulus s reflects the variability of behavioral responses we would find when repeatedly presenting the same stimulus s many times. This is conceptually distinct from the internal uncertainty of the observer on a single given trial, which is not directly measurable. Because of a strong prior, a Bayesian observer could have consistent responses from trial to trial despite being very internally uncertain on any particular trial. This completes the model: the distribution $p(\hat{s}|s)$ can now be compared to human behavior.

3.5. Case 5: Hierarchical Inference in Change Point Detection

Our last case is change point detection, the task of inferring from a time series of noisy observations whether or when an underlying variable changed. There are two main reasons for considering this task. First, it is a common form of inference. A new chef or owner might cause the quality of the food in a restaurant to suddenly change, or a neurologist may want to detect a seizure on an EEG in a comatose patient. Second, this case is representative of a type of inference that involves multiple layers of world state variables—in our case, not only the stimulus at each time point but also the "higher-level" variable of when the stimulus changed. In spite of this complication, the problem lends itself well to a Bayesian treatment (Wilson et al., 2013; Norton et al., 2019).

Step 1: Generative Model

The generative model is shown in Figure 8. Time is discrete and goes from 1 to T. A change occurs at exactly one time point t_{change}, chosen with equal probability:

$$p(t_{change}) = \frac{1}{T}. \quad \text{(Equation 13)}$$

The stimulus $\mathbf{s} = (s_1, \ldots, s_T)$ is a sequence that starts with repetitions of the value -1 and at some point changes to repetitions of the value 1. The change point is when the value -1 changes to 1. Formally,

$$s_t = \begin{cases} -1 & \text{if } t < t_{change} \\ 1 & \text{if } t \geq t_{change} \end{cases} \quad \text{(Equation 14)}$$

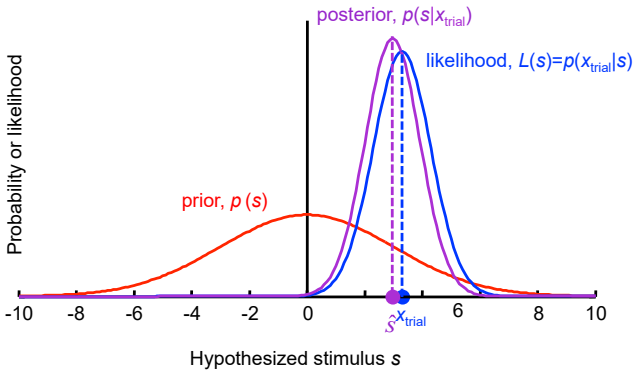


Figure 6. The Posterior Distribution Is Obtained by Multiplying the Prior by the Likelihood Function

Finally, we assume that the observer makes measurements $\mathbf{x} = (x_1,\ldots,x_T)$, whose noise is independent across time points. Mathematically, this means that we can write the conditional probability of the vector as the product of the conditional probabilities of the components:

$$p(\mathbf{x}\,|\,\mathbf{s}) = \prod_{t=1}^{T} p(x_t\,|\,s_t).$$

We next assume that each measurement follows a Gaussian distribution with mean equal to the stimulus at the corresponding time and with a fixed variance:

$$p(x_t\,|\,s_t) = \mathcal{N}(x_t; s_t, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_t - s_t)^2}{2\sigma^2}}.$$

Step 2: Inference

The observer would like to infer when the change occurred. What can make this problem challenging is that the noise in the measurements can create "apparent changes" that are not due to a change in the underlying state s_t. The Bayesian observer solves this problem in an optimal manner.

The world state of interest is change point t_{change}. Therefore, the Bayesian observer computes the posterior over t_{change} given a sequence of measurements, $\mathbf{x}$ (we leave out the subscript "obs" for ease of notation). We first apply Bayes' rule:

$$p(t_{\text{change}}|\mathbf{x}) \propto p(t_{\text{change}})p(\mathbf{x}\,|\,t_{\text{change}}). \qquad \text{(Equation 15)}$$

Since the prior is constant, $p(t_{\text{change}}) = \left(\frac{1}{T}\right)$, this simplifies to

$$p(t_{\text{change}}|\mathbf{x}) \propto p(\mathbf{x}\,|\,t_{\text{change}}). \qquad \text{(Equation 16)}$$

Thus, our challenge is to calculate the likelihood function

$$L(t_{\text{change}}) = p(\mathbf{x}\,|\,t_{\text{change}}). \qquad \text{(Equation 17)}$$

For each *hypothesized* value of t_{change}, this function tells us how *expected* the observations are under that hypothesis. The problem is that, unlike in case 4, the generative model does not give us this likelihood function right away; this is a consequence of the "stacked" or hierarchical nature of the generative model. We do have the distributions $p(\mathbf{x}\,|\,\mathbf{s})$ and $p(\mathbf{s}\,|\,t_{\text{change}})$. To make the link, we have to "average" over all possible values of $\mathbf{s}$. This is called "marginalization." Marginalization is extremely common in Bayesian models except for the very simplest ones, the reason being that there are almost always unknown states of the world that affect the observations but that the observer is not primarily interested in. The box describes the relevant probability calculus.

Marginalization of probabilities. If A and B are two discrete random variables, their "joint distribution" is $p(A,B)$. From the joint distribution, we can obtain the distribution of one of the variables by summing over the other one, for example:

$$p(A) = \sum_B p(A,B) \qquad \text{(Equation 18)}$$

This is called the "marginal distribution" of A. Making use of the definition of conditional probability, $p(A\,|\,B) = (p(A,B)\,/p(B))$, we further write

$$p(A) = \sum_B p(A\,|\,B)p(B). \qquad \text{(Equation 19)}$$

We can obtain a variant of this equation by conditioning each probability on a third random variable C:

$$p(A\,|\,C) = \sum_B p(A\,|\,B,C)p(B\,|\,C). \qquad \text{(Equation 20)}$$

This is the equation we use to obtain Equation 21.

We compute $L(t_{\text{change}})$ by marginalizing over $\mathbf{s}$:

$$L(t_{\text{change}}) = \sum_{\mathbf{s}} p(\mathbf{x}|\mathbf{s})p(\mathbf{s}\,|\,t_{\text{change}}), \qquad \text{(Equation 21)}$$

a. Besides using Equation 20, we used a property that is specific to our generative model. Which one?
b. For a given t_{change}, how many sequences $\mathbf{s}$ are possible?

Based on (b), we understand that the probability $p(\mathbf{s}\,|\,t_{\text{change}})$ is zero for all $\mathbf{s}$ except for one. Thus, the sum in Equation 21 reduces to a single term:

$$L(t_{\text{change}}) = p(\mathbf{x}\,|\,\text{the one}\,\mathbf{s}\,\text{in which the change occurs at}\,t_{\text{change}}) \qquad \text{(Equation 22)}$$

$$= \left(\prod_{t=1}^{t_{\text{change}}-1} p(x_t\,|\,s_t = -1)\right)\left(\prod_{t=t_{\text{change}}}^{T} p(x_t\,|\,s_t = 1)\right). \qquad \text{(Equation 23)}$$

This looks complicated, but we are not out of ideas!

c. Show that this can be written more simply as

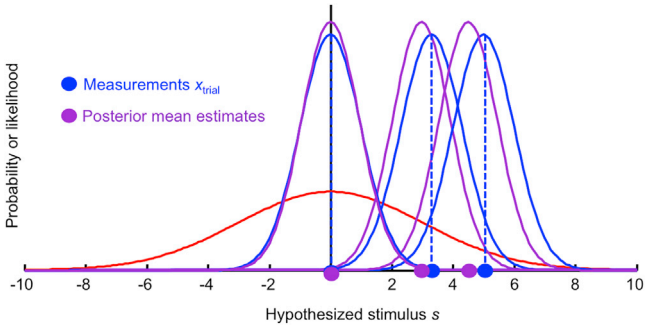


Figure 7. Trial-to-Trial Variability in Likelihoods and Posteriors
Likelihood functions (blue) and corresponding posterior distributions (purple) on three example trials on which the true stimulus *could have been the same*. The key point is that the likelihood function, the posterior distribution, and any posterior-derived estimate are not fixed objects: they move around from trial to trial because the measurement x_{trial} does.

$$L(t_{change}) \propto \prod_{t=t_{change}}^{T} \frac{p(x_t|s_t=1)}{p(x_t|s_t=-1)}. \qquad \text{(Equation 24)}$$

d. Now substitute $p(x_t|s_t) = \mathcal{N}(x_t; s_t, \sigma)$ to find

$$L(t_{change}) \propto e^{\frac{2}{\sigma^2}\sum_{t=t_{change}}^{T} x_t}. \qquad \text{(Equation 25)}$$

e. Does this equation make intuitive sense?

Combining Equations 16, 17, and 25, we find for the posterior probability of a change point at t_{change}:

$$p(t_{change}|\mathbf{x}) \propto e^{\frac{2}{\sigma^2}\sum_{t=t_{change}}^{T} x_t}. \qquad \text{(Equation 26)}$$

To obtain the actual posterior probabilities, the right-hand side has to be normalized (divided by the sum over all t_{change}).

f. The data in Figure 8 are $\mathbf{x} = (-0.46, 0.83, -3.26, -0.14, -0.68, -2.31, 0.57, 1.34, 4.58, 3.77)$, with $\sigma = 1$. Plot the posterior distribution over change point.

However, if the goal is just to pick the most probable change point (MAP estimate), normalizing is not needed.

g. Why not?
h. If that is the goal, the decision rule becomes "Cumulatively add up the elements of **x**, going backwards from $t = T$ to $t = 1$. The time at which this cumulative sum peaks is the MAP estimate of t_{change}." Explain.

Step 3: Response Distribution

In case 4, we were able to obtain an equation for the predicted response distribution. Here, however, and in many other cases, that is not possible. Nevertheless, we can still simulate the model's responses to obtain an approximate prediction.

i. Assume $\sigma = 1$ and $T = 10$. Vary the true change point t_{change} from 1 to T. For each value of t_{change}, we simulate 10,000 trials (or more if possible). On each simulated trial,
Based on t_{change}, specify the stimulus sequence **s**.
Simulate a measurement sequence **x** from **s**.
Apply the decision rule to each measurement sequence. The output is the simulated observer's response, namely an estimate of t_{change}.
Determine whether the response was correct.

Plot proportion correct as a function of t_{change}. Interpret the plot.

j. What is overall proportion correct (averaged across all t_{change})?
k. Vary $\sigma = 1, 2, 3$ and T from 2 to 16 in steps of 2. Plot overall proportion correct as a function of T for the three values of σ (color coded). Interpret the plot.
l. How could our simple example be extended to cover more realistic cases of change point detection?

4. Extensions

We have concluded our case studies of Bayesian decision models. These cases were basic, and many extensions can be made. We discuss a few such extensions here.

4.1. Multiple Observations

In case 4, if the decision maker makes two conditionally independent observations, x_1 and x_2, then the likelihood function $L(s)$ becomes the product $p(x_1|s)p(x_2|s)$. This is the premise of many Bayesian cue combination studies, in particular, multisensory ones (Trommershauser et al., 2011). If multiple observations are made over time, then Bayesian inference becomes a form of evidence accumulation. In case 5, multiple observations were also made over time, but, in addition, the stimulus changed as measurements were made. The most common generative model that describes time-varying stimuli with measurements at each time point is called a "Hidden Markov Model."

4.2. More Sophisticated Encoding Models

The distribution of an observation given a world state, $p(x|s)$, was deliberately kept very simple in cases 4 and 5. Instead, with x still being a noisy measurement of s, the noise level could depend on s, an instance of "heteroskedasticity"; this would make Bayesian inference substantially harder (Girshick et al., 2011; Acerbi et al., 2018). Furthermore, x could be a pattern of neural activity, for example, in early sensory cortex; this could be the starting point for asking how Bayesian computations could be implemented in neural populations (Ma et al., 2006).

4.3. More Realistic Cost Functions

Outside of purely perceptual tasks, an action is rarely an estimate of the state of the world. In fact, the action a can be of a very different nature than s. For example, if s represents whether milk is spoiled or still good, the action a could be to toss or drink the milk. Similarly, in an estimation experiment, a correct response could be rewarded differently depending on the true state of the world (Whiteley and Sahani, 2008). These scenarios can be captured by suitably choosing $C(a, s)$ in Equation 2.

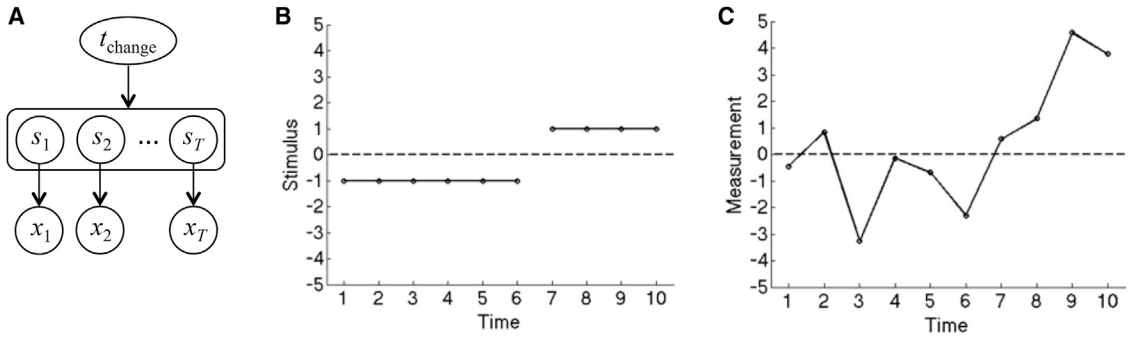


Figure 8. Case 5: Change Point Detection
(A) Generative model.
(B) Example of a true stimulus sequence with $t_{change} = 7$.
(C) Example sequence of measurements. When did the change from −1 to 1 occur?

4.4. Ad Hoc Sources of Error

Ad hoc sources of errors can be incorporated into Bayesian models as into any behavioral model. Those sources could include some proportion of random responses (lapse rate). In addition, the readout of the posterior could be stochastic rather than deterministic. In other words, decision noise could be added in step 2b. Such noise could alternatively reflect inevitable stochasticity, a level of imprecision that is strategically chosen to save effort when more effort would not be worth the task gains, or it could be a proxy for systematic but unmodeled suboptimalities in the decision process.

4.5. Model Mismatch

The decision maker might have wrong beliefs about the generative model. For example, if the true distributions of s and x are $p(s)$ and $p(x\,|\,s)$, the decision maker might instead believe that these variables follow different distributions, say $q(s)$ and $q(x\,|\,s)$. Then, they can still compute a posterior distribution $q(s\,|\,x_{trial})$, but it would be different from the correct one, $p(s\,|\,x_{trial})$. This is a case of "model mismatch": the distributions used in inference are not the same as in the true generative model (Beck et al., 2012).

5. Remarks

5.1. Persistent Myths

We address two common misunderstandings about Bayesian models of behavior. First, it is a myth that all Bayesian models are characterized by the presence of a non-uniform prior distribution. While the prior is important in many applications, the calculation of the likelihood is sometimes much more central. Examples include case 1 and most forms of sensory cue integration (Trommershauser et al., 2011). Second, it is a myth that all Bayesian models have so many degrees of freedom that any dataset can be fitted. On the contrary, as we saw in cases 4 and 5, Bayesian models have very few parameters if we assume that the observer uses the true generative model of the task. That being said, models with mismatch (Section 4.5) can have many more free parameters. However, adding parameters solely to obtain better fits violates the spirit of Bayesian modeling.

5.2. Bayesian versus Optimal Decision Making

Bayesian decision models are normative or optimal in the sense that, if the decision maker strives to minimize any form of total cost (or maximize any form of total reward) in the long run, then they should use the posterior distribution to determine their action. However, if the Bayesian decision maker suffers from model mismatch, then they are still Bayesian, but not necessarily optimal (Ma, 2012). In addition, a lot of recent work has focused on including in the decision maker's objective function not only task performance or task rewards, but also the cost of representation or computation. This gives rise to a class of modified Bayesian models known as "resource-rational" models (Griffiths et al., 2015).

6. Criticisms of Bayesian Models

Criticisms of Bayesian models of decision making fall into several broad categories. First, it has been alleged that Bayesian modelers insufficiently consider alternative models (Bowers and Davis, 2012). Indeed, many Bayesian modelers can do much better in testing Bayesian models against alternatives. This criticism, however, applies to many modeling studies, Bayesian or not. Second, it has been alleged that Bayesian models are overly flexible and can "fit anything" (Bowers and Davis, 2012). I consider this largely an unfair criticism, as Bayesian models are often highly constrained by either experimental or natural statistics (examples of the latter: Girshick et al., 2011; Geisler and Perry, 2009). It is true that some Bayesian studies use many parameters to agnostically estimate a prior distribution (e.g., Stocker and Simoncelli, 2006); those models need strong external validation (e.g., Houlsby et al., 2013) or to be appropriately penalized in model comparison. I discuss two more criticisms below and conclude by pointing out two major challenges for Bayesian models.

6.1. Bayesian Inference without and with Probabilities

It has been alleged that empirical findings cited in support of decision making being Bayesian are equally consistent with models that predict the same input-output mapping as the Bayesian model but without any representation of probabilities (Howe et al., 2006; Block, 2018). Indeed, in their weak form, Bayesian models are simply mappings from states to actions (policies) without regard to internal constructs such as likelihoods and posteriors (Ma and Jazayeri, 2014). In their strong form, however, Bayesian models claim that the brain represents those constructs. Such representations naturally give rise to notions of

uncertainty and decision confidence. For example, the standard deviation of a posterior distribution could be a measure of uncertainty. Evidence for the strong form could be obtained from Bayesian transfer tests (Maloney and Mamassian, 2009). Examples of Bayesian transfer tests involve varying sensory reliability (Trommershauser et al., 2011; Qamar et al., 2013), priors (Acerbi et al., 2014), or rewards (Whiteley and Sahani, 2008) from trial to trial without giving the subject performance feedback. However, when no transfer tests are done, it is difficult to argue that the brain has done more than learn a fixed policy that produces *as-if* Bayesian behavior. While Bayesian modelers need to be more explicit about the epistemological status of their models, and while some Bayesian studies only provide evidence for the weak form, the collective evidence for the strong form is by now plentiful (Ma and Jazayeri, 2014).

6.2. The Neural Implementation of Bayesian Inference

Bayesian modelers have been accused of not paying sufficient attention to the implementational level (Jones and Love, 2011). Bayesian models are primarily computational-level models of behavior. However, the evidence that decision-making behavior is approximately Bayesian in many tasks raises the question of how neurons implement Bayesian decisions. This question is most interesting for the strong form of Bayesian models because answering it then requires a theoretical commitment to the neural representations of likelihoods, priors, and cost functions. Consider the neural representation of a sensory likelihood function as an example. A straightforward theoretical postulate—but by no means the only one (Hoyer and Hyvärinen, 2003; Fiser et al., 2010; Deneve, 2008; Haefner et al., 2016)—is that the activity (e.g., firing rates) in a specific population of sensory neurons, collectively denoted by $\mathbf{r}$, takes the role of the observation, so that the sensory component of the generative model—the analog of Equation 5—becomes a stimulus-conditioned activity distribution $p(\mathbf{r}\,|\,s)$ (Sanger, 1996; Pouget et al., 2003). One could then proceed by hypothesizing that this distribution serves as the basis of the likelihood function. In other words, for given activity $\mathbf{r}_{\text{trial}}$, the likelihood of stimulus s would be

$$L(s) = p(\mathbf{r}_{\text{trial}}\,|\,s), \qquad \text{(Equation 27)}$$

in analogy to Equation 6. This equation is the cornerstone of the theory of ''probabilistic population coding'' (Ma et al., 2006). The hypothesis would be untestable if it were not for the brain's use of a likelihood function over s in subsequent computation. Therefore, one needs a behavioral task in which evidence for the strong form of a Bayesian model has been obtained. Then, one could use Equation 27 to decode on each trial a full neural likelihood function from a sensory population and plug it into the Bayesian model to predict behavior (van Bergen et al., 2015; Walker et al., 2019).

Equipped with putative neural representations of the building blocks of Bayesian inference, one can ask the further question of how the computation that puts the pieces together is implemented. In a handful of cases, this question can be approached analytically. For example, in cue combination, the Bayesian computation consists of a multiplication of likelihood functions. Under a certain assumption about $p(\mathbf{r}_{\text{trial}}\,|\,s)$, the neural implementation of this computation is a simple addition of patterns of neural activity (Ma et al., 2006). However, this case might be an exception because, under the same distributional assumption, the neural implementation of many other Bayesian computations is not exact and can get very complex (e.g., Ma et al., 2011). Instead, a simple trained neural network can perform strong-form Bayesian inference across a wide variety of tasks, even when the distributional assumption is violated (Orhan and Ma, 2017). The neural implementation of Bayesian computation remains an active area of research.

Finally, resource-rational theories take implementation-level costs and constraints seriously (Griffiths et al., 2015). A resource-rational decision maker is one who not only maximizes task performance, but also simultaneously minimizes an ecologically meaningful cost, such as total firing rate, total number of neurons, amount of effort, or time spent. Resource-rational models provide accounts of the nature of representations as well as of apparent suboptimalities in decision making.

6.3. Other Challenges

We briefly mention two other major challenges to Bayesian models of decision making. The first is that they often do not scale up well. For example, if one were to infer the depth ordering of N image patches, there are $N!$ possible orderings, and a Bayesian decision maker would have to consider every one of them. For large N, this would be computationally prohibitive and not realistic as a model of human vision. A similar combinatorial explosion would arise in case 5 if the number of change points were not known to be 1. A second challenge is how a Bayesian decision maker learns a natural generative model from scratch using only a small number of training examples (Tenenbaum et al., 2011; Lake et al., 2015).

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.neuron.2019.09.037.

ACKNOWLEDGMENTS

This primer is based on a Bayesian modeling tutorial that I have taught in several places. Thanks to all students who actively participated in these tutorials. Special thanks to my teaching assistants of the Bayesian tutorial at the Computational and Systems Neuroscience conference in 2019, who not only taught but also greatly improved the five case studies and wrote solutions: Anna Kutschireiter, Anne-Lene Sax, Jennifer Laura Lee, Jorge Menéndez, Julie Lee, Lucy Lai, and Sashank Pisupati. A much more detailed didactic introduction to Bayesian decision models will appear in 2020 in book form; many thanks to my co-authors of that book, Konrad Körding and Daniel Goldreich. My research is funded by grants R01EY020958, R01EY027925, R01MH118925, and R01EY026927 from the National Institutes of Health.

REFERENCES

Acerbi, L., Vijayakumar, S., and Wolpert, D.M. (2014). On the origins of suboptimality in human probabilistic inference. PLoS Comput. Biol. *10*, e1003661.

Acerbi, L., Dokka, K., Angelaki, D.E., and Ma, W.J. (2018). Bayesian comparison of explicit and implicit causal inference strategies in multisensory heading perception. PLoS Comput. Biol. *14*, e1006110.

Battaglia, P.W., Hamrick, J.B., and Tenenbaum, J.B. (2013). Simulation as an engine of physical scene understanding. Proc. Natl. Acad. Sci. USA *110*, 18327–18332.

Beck, J.M., Ma, W.J., Pitkow, X., Latham, P.E., and Pouget, A. (2012). Not noisy, just wrong: the role of suboptimal inference in behavioral variability. Neuron *74*, 30–39.

Block, N. (2018). If perception is probabilistic, why does it not seem probabilistic? Philos. Trans. R. Soc. Lond. B Biol. Sci. *373*, 20170341.

Bowers, J.S., and Davis, C.J. (2012). Bayesian just-so stories in psychology and neuroscience. Psychol. Bull. *138*, 389–414.

Brainard, D.H., and Freeman, W.T. (1997). Bayesian color constancy. J. Opt. Soc. Am. A Opt. Image Sci. Vis. *14*, 1393–1411.

Cogley, T., and Sargent, T.J. (2008). Anticipated utility and rational expectations as approximations of Bayesian decision making. Int. Econ. Rev. *49*, 185–221.

Deneve, S. (2008). Bayesian spiking neurons I: inference. Neural Comput. *20*, 91–117.

Faisal, A.A., Selen, L.P., and Wolpert, D.M. (2008). Noise in the nervous system. Nat. Rev. Neurosci. *9*, 292–303.

Fiser, J., Berkes, P., Orbán, G., and Lengyel, M. (2010). Statistically optimal perception and learning: from behavior to neural representations. Trends Cogn. Sci. *14*, 119–130.

Geisler, W.S., and Perry, J.S. (2009). Contour statistics in natural images: grouping across occlusions. Vis. Neurosci. *26*, 109–121.

Girshick, A.R., Landy, M.S., and Simoncelli, E.P. (2011). Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. Nat. Neurosci. *14*, 926–932.

Goodman, N.D., and Frank, M.C. (2016). Pragmatic language interpretation as probabilistic inference. Trends Cogn. Sci. *20*, 818–829.

Griffiths, T.L., Lieder, F., and Goodman, N.D. (2015). Rational use of cognitive resources: levels of analysis between the computational and the algorithmic. Top. Cogn. Sci. *7*, 217–229.

Haefner, R.M., Berkes, P., and Fiser, J. (2016). Perceptual decision-making as probabilistic inference by neural sampling. Neuron *90*, 649–660.

Houlsby, N.M., Huszár, F., Ghassemi, M.M., Orbán, G., Wolpert, D.M., and Lengyel, M. (2013). Cognitive tomography reveals complex, task-independent mental representations. Curr. Biol. *23*, 2169–2175.

Howe, C.Q., Beau Lotto, R., and Purves, D. (2006). Comparison of Bayesian and empirical ranking approaches to visual perception. J. Theor. Biol. *241*, 866–875.

Hoyer, P.O., and Hyvärinen, A. (2003). Interpreting neural response variability as Monte Carlo sampling of the posterior. In Advances in Neural Information Processing Systems (NIPS), 293–30.

Itti, L., and Baldi, P. (2009). Bayesian surprise attracts human attention. Vision Res. *49*, 1295–1306.

Jones, M., and Love, B.C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. Behav. Brain Sci. *34*, 169–188.

Kersten, D., Mamassian, P., and Yuille, A. (2004). Object perception as Bayesian inference. Annu. Rev. Psychol. *55*, 271–304.

Körding, K.P., and Wolpert, D.M. (2004). Bayesian integration in sensorimotor learning. Nature *427*, 244–247.

Lake, B.M., Salakhutdinov, R., and Tenenbaum, J.B. (2015). Human-level concept learning through probabilistic program induction. Science *350*, 1332–1338.

Ma, W.J. (2012). Organizing probabilistic models of perception. Trends Cogn. Sci. *16*, 511–518.

Ma, W.J., and Jazayeri, M. (2014). Neural coding of uncertainty and probability. Annu. Rev. Neurosci. *37*, 205–220.

Ma, W.J., Beck, J.M., Latham, P.E., and Pouget, A. (2006). Bayesian inference with probabilistic population codes. Nat. Neurosci. *9*, 1432–1438.

Ma, W.J., Navalpakkam, V., Beck, J.M., Berg, Rv., and Pouget, A. (2011). Behavior and neural basis of near-optimal visual search. Nat. Neurosci. *14*, 783–790.

Maloney, L.T., and Mamassian, P. (2009). Bayesian decision theory as a model of human visual perception: testing Bayesian transfer. Vis. Neurosci. *26*, 147–155.

Martí, R., Lozano, J.A., Mendiburu, A., and Hernando, L. (2016). Multi-start methods. In Handbook of Heuristics, R. Martí, P. Pardalos, and M. Resende, eds. (Springer), pp. 155–175.

Myung, I.J. (2003). Tutorial on maximum likelihood estimation. J. Math. Psychol. *47*, 90–100.

Norton, E.H., Acerbi, L., Ma, W.J., and Landy, M.S. (2019). Human online adaptation to changes in prior probability. PLoS Comput. Biol. *15*, e1006681.

Orhan, A.E., and Ma, W.J. (2017). Efficient probabilistic inference in generic neural networks trained with non-probabilistic feedback. Nat. Commun. *8*, 138.

Pouget, A., Dayan, P., and Zemel, R.S. (2003). Inference and computation with population codes. Annu. Rev. Neurosci. *26*, 381–410.

Qamar, A.T., Cotton, R.J., George, R.G., Beck, J.M., Prezhdo, E., Laudano, A., Tolias, A.S., and Ma, W.J. (2013). Trial-to-trial, uncertainty-based adjustment of decision boundaries in visual categorization. PNAS *110*, 20332–20337.

Sanger, T.D. (1996). Probability density estimation for the interpretation of neural population codes. J. Neurophysiol. *76*, 2790–2793.

Stocker, A.A., and Simoncelli, E.P. (2006). Noise characteristics and prior expectations in human visual speed perception. Nat. Neurosci. *9*, 578–585.

Tenenbaum, J.B., Griffiths, T.L., and Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. Trends Cogn. Sci. *10*, 309–318.

Tenenbaum, J.B., Kemp, C., Griffiths, T.L., and Goodman, N.D. (2011). How to grow a mind: statistics, structure, and abstraction. Science *331*, 1279–1285.

Treisman, M. (1977). Motion sickness: an evolutionary hypothesis. Science *197*, 493–495.

Trommershauser, J., Kording, K., and Landy, M.S. (2011). Sensory Cue Integration (Oxford University Press).

van Bergen, R.S., Ma, W.J., Pratte, M.S., and Jehee, J.F. (2015). Sensory uncertainty decoded from visual cortex predicts behavior. Nat. Neurosci. *18*, 1728–1730.

Wagemans, J., Feldman, J., Gepshtein, S., Kimchi, R., Pomerantz, J.R., van der Helm, P.A., and van Leeuwen, C. (2012). A century of Gestalt psychology in visual perception: II. Conceptual and theoretical foundations. Psychol. Bull. *138*, 1218–1252.

Walker, E.Y., Cotton, R.J., Ma, W.J., and Tolias, A.S. (2019). A neural basis of probabilistic computation in visual cortex. bioRxiv. https://doi.org/10.1101/365973.

Wertheimer, M. (1938). Gestalt Theory. In A Source Book of Gestalt Psychology, W.D. Ellis, ed. (Kegan Paul, Trench, Trubner & Company), pp. 1–11.

Whiteley, L., and Sahani, M. (2008). Implicit knowledge of visual uncertainty guides decisions with asymmetric outcomes. J. Vis. *8*, 1–15.

Wilson, R.C., Nassar, M.R., and Gold, J.I. (2013). A mixture of delta-rules approximation to bayesian inference in change-point problems. PLoS Comput. Biol. *9*, e1003150.

Neuron, Volume *104*

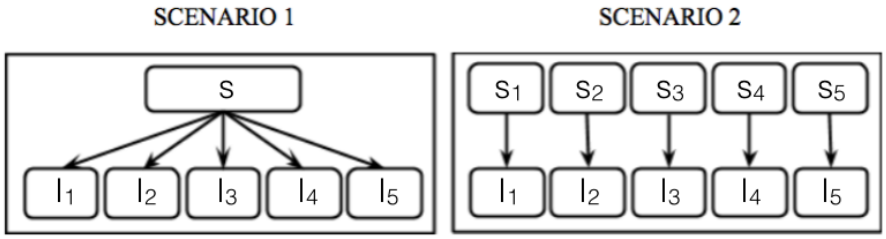
Supplemental Information

Bayesian Decision Models: A Primer

Wei Ji Ma

A Solutions to exercises

A.1 Case 1: Unequal likelihoods and Gestalt laws



a.

b. $\frac{1}{2}$

c. $\frac{1}{32}$

d. The likelihoods do not add up to 1, nor should they. How probable the observations are under one scenario about the world is independent of how probable those same observations are under another scenario about the world.

e. The phrase is misleading because it suggests that there is only one. There are multiple different likelihoods for the same set of observations, one for each hypothetisized state of the world.

f. $p(H_1) = \frac{2}{3}$, $p(H_2) = \frac{1}{3}$

g. $\frac{1}{3}$

h. $\frac{1}{96}$

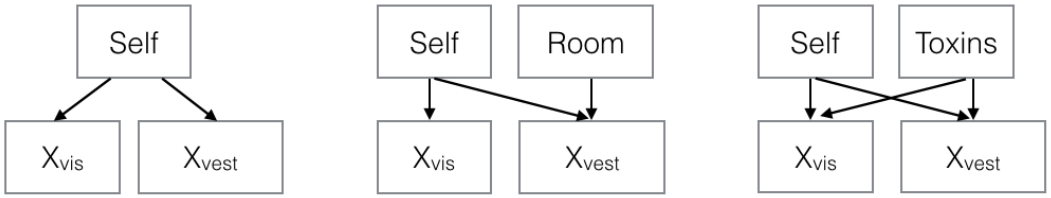
i. No.

j. $P(H_1|I_1, \ldots, I_5) = 0.97$, $P(H_2|I_1, \ldots, I_5) = 0.03$.

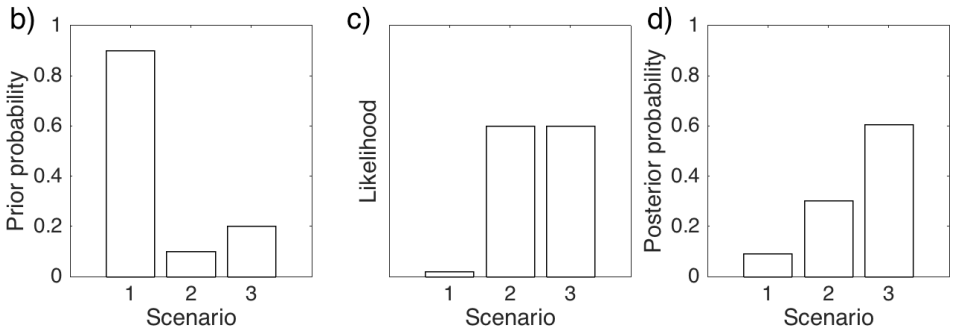
k. Since H_1 has the highest posterior probability, we are expected to perceive H_1– i.e., we perceive the group of dots as being part of the same object. This is consistent with the Gestalt law of common fate.

l. The Bayesian account produces predictions consistent with the Gestalt law but it goes beyond the "descriptive" law by providing an account that might be considered both normative and perhaps explanatory– it tells us what the optimal observer should perceive, beyond merely describing what people do perceive. Paired with certain evolutionary premises, the Bayesian account might provide an *explanation* of the law of common fate.

A.2 Case 2: Competing likelihoods and priors in motion sickness



a.



e. E.g. priors $= (0.9, 0.1, 0.2)$, likelihoods $= (0.01, 0.3, 0.3)$, posteriors $= (0.09, 0.3, 0.61)$.
f. Our evolutionary priors tell us stationary rooms are very probable, moving rooms are highly improbable, and the ingestion of a toxin is somewhat rare. The likelihood information we receive as a result of there being a visual-vestibular mismatch suggests that the stationary room hypothesis is improbable, and each of the other two hypotheses (which very likely lead to sensory mismatches) are probable. The posteriors are found by multiplying the prior and likelihood for each hypothesis. Scenario 3, the hypothesis that you are hallucinating because you've ingested a toxin, yields the highest posterior probability. If your body believes the MAP hypothesis that a toxin was ingested, it might trigger vomiting as a natural defensive response.

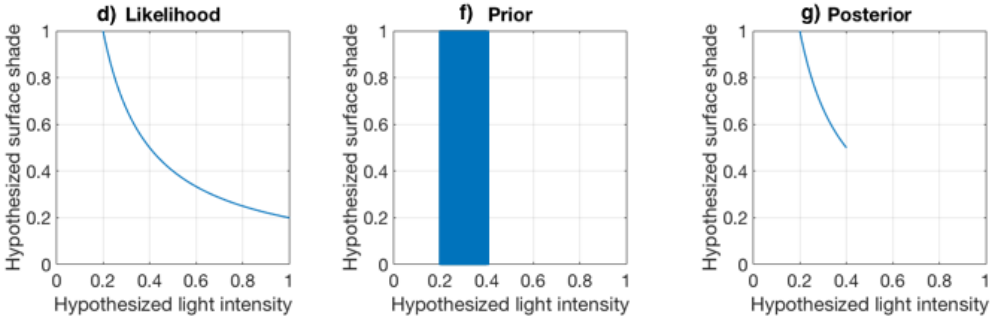
A.3 Case 3: Ambiguity from a nuisance parameter: Surface shade perception

a. 0.2

b. 0.5

c. Intuitively, the same retinal intensity might be caused by a bright light source and a dark surface , or a dark light source and light surface. Retinal intensity alone provides ambiguous information about surface shade, since we do not know the contribution of light intensity.

d.



Each plot should be interpreted as a heat map (with white background = 0).

e. This graph shows the likelihoods of each hypothesized combination of light intensity and surface shade, when the observed retinal intensity is 0.2. World states which lie off of this curve cannot produce a retinal intensity of 0.2. World states on this curve definitely produce a retinal intensity of 0.2.

f. To adjudicate between all possible points along this curve, the visual system might take prior over light intensity into account. For instance, if it is dark outside, there would be high prior probabilities for low light intensities. Combining with the likelihood, the posterior might be highest for the hypothesis that the surface of the object is white.

g. See figure.

h. See figure.

i. We would perceive the surface as having a shade somewhere between 0.5 and 1.

A.4 Case 4: Inference under measurement noise in sound localization

a. The proportionality sign is appropriate because s-independent factors are irrelevant when performing the final normalizing step needed to obtain the posterior.

b. The following piece of Matlab code approximately reproduces the figure:

```
clear, close all;
svec = -10:0.01:10;
mu = 0;
sig_s = 2;
sig = 1;
x_trial = 3.2;
prior = normpdf(svec, mu, sig_s);
prior = prior / sum(prior);
likelihood = normpdf(svec, x_trial, sig);
likelihood = likelihood / sum(likelihood);
protoposterior = prior .* likelihood;
posterior = protoposterior / sum(protoposterior);
figure; hold on
plot(svec,prior,'r')
plot(svec,likelihood,'b')
plot(svec,posterior,'k')
```

c. Posterior:

$$p(s|x_{\text{trial}}) \propto p(s)p(x_{\text{trial}}|s) \propto e^{-\frac{1}{2}\left(\frac{(s-\mu)^2}{\sigma_s^2}+\frac{(x_{\text{trial}}-s)^2}{\sigma^2}\right)}$$

We use the notation $J_s = \frac{1}{\sigma_s^2}$ and $J = \frac{1}{\sigma^2}$. Using the hint, we rewrite as

$$p(s|x_{\text{trial}}) \propto e^{-\frac{1}{2}\left[(J_s+J)\left(s-\frac{\mu J_s + x_{\text{trial}} J}{J_s+J}\right)^2 + \text{junk}\right]}$$
$$\propto e^{-\frac{\left(s-\frac{\mu J_s + x_{\text{trial}} J}{J_s+J}\right)^2}{2\frac{1}{J_s+J}}}$$

This is the equation for a Gaussian with mean and variance as given in the problem.

d. $\frac{\mu + x_{\text{trial}}}{2}$.

e. Should be straightforward.

f. Using the result of (e), $\sigma^2_{\text{posterior}} < \sigma^2$ because $\frac{\sigma_s^2}{\sigma^2+\sigma_s^2} < 1$. Analogously, $\sigma^2_{\text{posterior}} < \sigma_s^2$.

g. $\frac{\sigma^2}{2}$

h. As a rule for normal distributions, if $y \sim \mathcal{N}(\mu, \sigma^2)$, then $ay+b \sim \mathcal{N}(a\mu+b, a^2\sigma^2)$, e.g. scaling a Gaussian distribution by a factor of 2 would increase its variance by a factor of 4. The posterior mean estimate $\hat{s} = \mu_{\text{posterior}} = wx + (1-w)\mu$ involves scaling the measurement x by a constant w and shifting it by constant $(1-w)\mu$, where $w = \frac{\frac{1}{\sigma^2}}{\frac{1}{\sigma^2}+\frac{1}{\sigma_s^2}}$. Therefore, the estimate distribution will have a variance which scales the variance of the measurement distribution by w^2:

$$p(x|s) \sim N(x; s, \sigma^2)$$
$$p(\hat{s}|s) \sim N(\hat{s}; ws + (1-w)\mu, w^2\sigma^2)$$
$$\text{Var}(\hat{s}|s) = w^2\sigma^2 = \left(\frac{\frac{1}{\sigma^2}}{\frac{1}{\sigma^2}+\frac{1}{\sigma_s^2}}\right)^2 \sigma^2 = \frac{\frac{1}{\sigma^2}}{(\frac{1}{\sigma^2}+\frac{1}{\sigma_s^2})^2}$$

A.5 Case 5: Hierarchical inference in change point detection

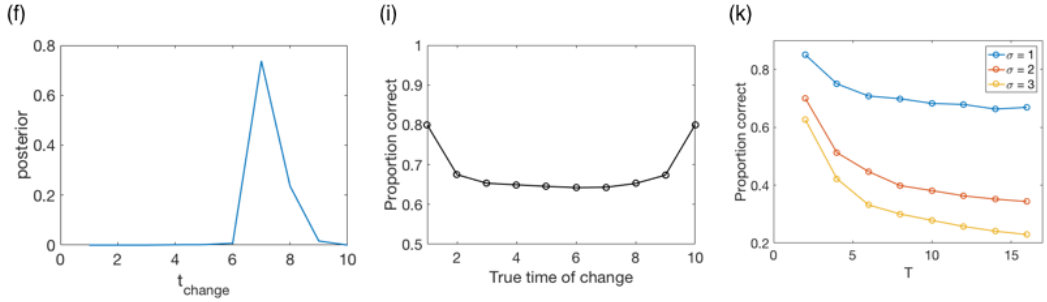
a. The property that once the sequence $\mathbf{s}$ is given, $\mathbf{x}$ is independent of t_{change}. In other words, $p(\mathbf{x}|\mathbf{s}, t_{\text{change}}) = p(\mathbf{x}|\mathbf{s})$.

b. For a given t_{change}, only one sequence is possible.

c. Divide by the constant $\prod_{t=1}^{T} p(x_t|s_t = -1)$. This is a constant because it does not depend on t_{change}, and it can therefore be absorbed into the proportionality sign.

d.

$$L(t_{\text{change}}) \propto \prod_{t=t_{\text{change}}}^{T} \frac{p(x_t|s_t=1)}{p(x_t|s_t=-1)} = \prod_{t=t_{\text{change}}}^{T} e^{-\frac{(x_t-1)^2}{2\sigma^2}+\frac{(x_t+1)^2}{2\sigma^2}} = e^{\frac{2}{\sigma^2}\sum_{t=t_{\text{change}}}^{T} x_t}$$



g. Normalizing does not change the most probable change point, since it is just the highest point on the graph.

j. 0.683.

l. The change point could have a higher probability of occurring at some times rather than others. There could be multiple change points instead of one. The stimulus could take more values than merely -1 or 1.