Forthcoming, Handbook of Perceptual Organization (J. Wagemans, Ed.)

Bayesian models of perception: a tutorial introduction

Jacob Feldman
Dept. of Psychology, Center for Cognitive Science
Rutgers University - New Brunswick

Abstract

Bayesian approaches to perception offer a principled, coherent and elegant answer to the central problem of perception: what the brain should believe about the world based on sensory data. This chapter gives a tutorial introduction to Bayesian inference, illustrating how it has been applied to problems in perception.

Inference in perception

One of the central ideas in the study of perception is that the proximal stimulus—the pattern of energy that impinges on sensory receptors, such as the visual image—is not sufficient to specify the actual state of the world outside (the distal stimulus). That is, while the image of your grandmother on your retina might look like your grandmother, it also looks like an infinity of other arrangements of matter, each having a different combination of 3D structure, surface properties, color properties, etc., so that they happen to look just like your grandmother from a particular viewpoint. Naturally, the brain generally does not perceive these far-fetched alternatives, but rapidly converges on a single solution which is what we consciously perceive. A shape on the retina might be a large object that is far away, or a smaller one more nearby, or anything in between. A mid-gray region on the retina might be a bright white object in dim light, or a dark object in bright light, or anything in between. An elliptical shape on the retina might be an elliptical object face-on, or a circular object slanted back in depth, or anything in between. Every proximal stimulus is consistent with an infinite family of possible scenes, only one of which is perceived.

The central problem for the perceptual system is to quickly and reliably decide among all these alternatives, and the central problem for visual science is to figure out what rules, principles, or mechanisms the brain uses to do so. This process was called *unconscious inference* by Helmholtz, perhaps the first scientist to appreciate the problem, and is sometimes called *inverse optics* to convey the idea that the brain must in a sense

I am grateful to Manish Singh, Vicky Froyen, and Johan Wagemans for helpful discussions. Preparation of this article was supported by NIH EY021494. Please direct correspondence to the author at jacob@ruccs.rutgers.edu.

invert the process of optical projection—to take the image and recover the world that gave rise to it.

The modern history of visual science contains a wealth of proposals for how exactly this process works, far too numerous to review here. Some are very broad, like the Gestalt idea of *Prägnanz* (infer the simplest or most reasonable scene consistent with the image). Many others are narrowly addressed to specific aspects of the problem like the inference of shape or surface color. But historically, the vast majority of these proposals suffer from one (or both) of the following two problems. First, many (like Prägnanz and many other older suggestions) are too vague to be realized as computational mechanisms. They rest on central ideas, like the Gestalt term "goodness of form," that are at best can only be subjectively defined and cannot be implemented algorithmically without a host of additional assumptions. Second, many proposed rules are arbitrary or unmotivated, meaning that is unclear exactly why the brain would choose them rather than an infinity of other equally effective ones. Of course, it cannot be taken for granted that mental processes are principled in this sense, and some have argued for a view of the brain as a "bag of tricks" (Ramachandran, 1985). Nevertheless, to many theorists, a mental function as central and evolutionarily ancient as perceptual inference seems to demand a more coherent and principled explanation.

Inverse probability and Bayes' rule

In recent decades, Bayesian inference has been proposed as a solution to these problems, representing a principled, mathematically well-defined, and comprehensive solution to the problem of inferring the most plausible interpretation of sensory data. Bayesian inference begins with the mathematical notion of *conditional probability*, which is simply probability restricted to some particular set of circumstances. For example, the conditional probability of A conditioned on B, denoted $p(A|B)$, means the probability that A is true given that B is true. Mathematically, this conditional probability is simply the ratio of the probability of that A and B are both true, $p(A \text{ and } B)$, divided by the probability that B is true, $p(B)$, hence

$$p(A|B) = \frac{p(A \text{ and } B)}{p(B)}. \tag{1}$$

Similarly, the probability of B given A is the ratio of the probability that B and A are both true divided by the probability that A is true, hence

$$p(B|A) = \frac{p(B \text{ and } A)}{p(A)}. \tag{2}$$

It was the reverend Thomas Bayes (1763) who first noticed that these mathematically simple observations can be combined[1] to yield a formula for the conditional probability $p(A|B)$ (A given B) in terms of the *inverse* conditional probability $p(B|A)$ (B given A),

[1]More specifically, note that $p(B \text{ and } A) = p(A \text{ and } B)$ (conjunction is commutative). Substitute the latter for the former in Eq. 1 to see that $p(A|B)p(B)$, and likewise $p(B)p(A|B)$, are both equal to $p(A \text{ and } B)$ and thus to each other. Divide both sides of $p(A|B)p(B) = p(B|A)p(A)$ by $p(B)$ to yield Bayes' rule.

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}, \tag{3}$$

a formula now called *Bayes' theorem* or *Bayes' rule*.[2] Before Bayes, the mathematics of probability had been used exclusively to calculate the chances of a particular random outcome of a stochastic process, like the chance of getting ten consecutive heads in ten flips of a fair coin [p(10 heads|fair coin)]. Bayes realized that his rule allowed us to *invert* this inference and calculate the probability of the conditions that gave rise to the observed outcome—here, the probability, having observed 10 consecutive heads, that the coin was fair in the first place [p(fair coin|10 heads)]. Of course, to determine this, you need to assume that there is some *other* hypothesis we might entertain about the state of the coin, such as that it is biased towards heads. Bayes' logic, often called *inverse probability*, allows us to evaluate the plausibility of various hypotheses about the state of the world (the nature of the coin) on the basis of what we have observed (the sequence of flips). For example, it allows us to quantify the degree to which observing 10 heads in a row might persuade us that the coin is biased towards heads.

Bayes and his followers, especially the visionary French mathematician Laplace, saw how inverse probability could form the basis of a full-fledged theory of inductive inference (see Stigler, 1986). As David Hume had pointed out only a few decades previously, much of what we believe in real life—including all generalizations from experience—cannot be proved with logical certainty, but instead merely seems intuitively plausible on the basis of our knowledge and observations. To philosophers seeking a deductive basis for our beliefs, this argument was devastating. But Laplace realized that Bayes' rule allowed us to *quantify* belief—to precisely gauge the plausibility of inductive hypotheses.

By Bayes' rule, given any data D which has a variety of possible hypothetical causes H_1, H_2, etc., each cause H_i is plausible in proportion to the product of two numbers: the probability of the data if the hypothesis is true $p(D|H_i)$, called the *likelihood*; and the *prior probability* of the hypothesis, $p(H_i)$, that is, how probable the hypothesis was in the first place. If the various hypotheses are all mutually exclusive, then the probability of the data D is the sum of its probability under all the various hypotheses,

$$p(D) = p(H_1)p(D|H_1) + p(H_1)p(D|H_1) + \ldots = \sum_i p(H_i)p(D|H_i). \tag{4}$$

Plugging this into Bayes' rule (with H_i playing the role of A, and D playing the role of B), this means that the probability of hypothesis H_i given data D, called the *posterior probability* $p(H_i|D)$, is

$$p(H_i|D) = \frac{p(H_i)p(D|H_i)}{p(D)} = \frac{p(H_i)p(D|H_i)}{\sum_i p(H_i)p(D|H_i)}, \tag{5}$$

or in words:

$$\text{posterior for } H_i = \frac{\text{prior for } H_i \times \text{likelihood of } H_i}{\text{sum of (prior } \times \text{likelihood) over all hypotheses}}. \tag{6}$$

[2] Actually, the rule does not appear in this form in Bayes' essay. But Bayes' focus was indeed on the underlying problem of inverse inference and deserves credit for the main insight. See Stigler (1983).

The posterior probability $p(H_i|D)$ quantifies how much we should believe H_i after considering the data. It is simply the ratio of the probability of the evidence under H_i (the product of its prior and likelihood) relative to the *total* probability of the evidence arising under all hypotheses (the sum of the prior-likelihood products for all the hypotheses). This ratio measures how plausible H_i is relative to all the other hypotheses under consideration.

But Laplace's ambitious account was followed by a century of intense controversy about the use of inverse probability (see Howie, 2004). In modern retellings, critics' objection to Bayesian inference is often reduced to the idea that to use Bayes' rule we need to know the prior probability of each of the hypotheses (for example, the probability the coin was fair in the first place), and that we often don't have this information. But their criticism was far more fundamental and relates to the meaning of probability itself. They argued that many propositions—those that refer to propositions whose truth value is fixed though unknown—can't be assigned probabilities at all, in which case the use of inverse probability would be nonsensical. This criticism reflects a conception of probability, often called *frequentism*, in which probability refers exclusively to *relative frequency* in a repeatable chance situation. Thus, in their view, you can calculate the probability of a string of heads for a fair coin, because this is a random event that occurs on some fraction of trials; but you can't calculate a probability of a of a *non*-repeatable state of nature, like *this coin is fair*, or *the Higgs boson exists* because such hypotheses are either definitely true or definitely false, and are not "random." The frequentist objection was not just that we don't know the prior for many hypotheses, but that most hypotheses don't *have* priors—or posteriors, or any probabilities at all.

But in contrast, Bayesians generally thought of probability as quantifying the *degree of belief*, and were perfectly content to apply it to any proposition at all, including non-repeatable ones. To Bayesians, the probability of any proposition is simply a characterization of our state of knowledge about it, and can freely be applied to *any* proposition as a way of quantifying how strongly we believe it. This conception of probability, sometimes called *subjectivist* (or *epistemic* or sometimes just *Bayesian)*, is thus essential to the Bayesian program. Without it, one cannot calculate the posterior probability of a non-repeatable proposition, because such propositions simply don't have probabilities—and this would rule out most uses of Bayes' rule to perform induction.

The sometimes ferocious controversy over this issue culminated around 1920 when the fervently frequentist statisticians Fisher, Neyman, and Pearson founded what we now call classical statistics—sampling distributions, significance tests, confidence intervals, and so forth—on a platform of rejecting inverse probability in the name of objectivity. But the theory of Bayesian inference continued to develop in the shadows, and was given a comprehensive modern formulation by Harold Jeffreys (1939/1961) and others. This history helps explain why, despite centuries of development, Bayesian techniques are only in the last few decades being applied without apology to inference problems in many fields, including human cognition.

Bayesian inference as a rational model of perception

The development of Bayesian theory in the 20th century was invigorated by the discovery of something quite remarkable about Bayesian inference: it is *rational,* and uniquely so. Cox (1961) showed that Bayesian inference has the unique property that it,

and it alone among inference systems, satisfies basic considerations of internal consistency, such as invariance to the order in which evidence is considered. If one wishes to assign degrees of belief to hypotheses in a rational way, one must inevitably use the conventional rules of probability, and specifically Bayes' rule. Later de Finetti (see de Finetti, 1970/1974) demonstrated the uniquely rational status of Bayesian inference in an even more acute way. He showed that if a system of inference differs from Bayesian inference in any substantive way, it is subject to catastrophic failures of rationality. (His so-called Dutch book theorem shows, in essence, that any non-Bayesian reasoner can be turned into a "money pump".) In recent decades these strong arguments for Bayesian inference as a uniquely rational system for fixing belief were brought to wide attention by the vigorous advocacy of the physicist E. T. Jaynes (see Jaynes, 2003). Though there are of course many subtleties surrounding the supposedly optimal nature of Bayesian inference (see Earman, 1992), most contemporary statisticians have rejected the dogmatic frequentism that underlies classical statistics, and now regard Bayesian inference as an optimal method for making inferences on the basis of data.

This characterization of Bayesian inference—as an optimal method for deciding what to believe under conditions of uncertainty—makes it perfectly suited to the central problem of perception, that of estimating the properties of the physical world based on sense data. The basic idea is to think of the stimulus (e.g. the visual image) as reflecting both stable properties of the world (which we would like to infer) plus some uncertainty introduced in the process of image formation (which we would like to disregard). Bayesian inference allows us estimate the stable properties of the world conditioned on the image data. The aptness of Bayesian inference as a model of perceptual inference was first noticed in the 1980s by a number of authors, and brought to wider attention by the collection of papers in Knill and Richards (1996). Since then the applications of Bayes to perception have multiplied and evolved, while always retaining the core idea of associating perceptual belief with the posterior probability as given by Bayes' rule. Several excellent reviews of the literature are already available (e.g. see Kersten, Mamassian, & Yuille, 2004; Knill, Kersten, & Yuille, 1996; Yuille & Bülthoff, 1996) each with a slightly different emphasis or slant. The current chapter is intended to be at a tutorial introduction to the main ideas of Bayesian inference in human perception, with some emphasis on misunderstandings that tend to arise in the minds of newcomers to the topic. Although the examples are drawn from the perception literature, most of the main ideas apply equally to other areas of cognition as well. The emphasis will be on central principles rather than on mathematical details or recent technical advances.

Basic calculations in Bayesian inference

We begin with several simple numerical examples to illustrate the basic calculations in Bayesian inference, before moving on to perceptual examples.

Bayesian inference for discrete hypotheses

The simplest type of Bayesian inference involves a finite number of distinct hypotheses $H_1 \ldots H_n$, each of which has a prior probability $p(H_i)$ and a likelihood function $p(X|H_i)$

which gives the probability of each possible dataset X conditioned on that hypothesis.[3] For example, imagine that you hear a noise X on your roof, which is either an animal A or a burglar B. The noise sounds a bit like an animal, implying a moderate animal likelihood, say $p(X|A) = .3$. (That is, if it *were* an animal, there is about a 30% chance of a noise of the type that you hear.) But unfortunately it sounds a lot like a burglar, implying a high burglar likelihood, say $p(X|B) = .8$. Classical statistics dictates that we select hypotheses by maximizing likelihood, which in this situation would imply a burglar (and necessitate an immediate call to the police). But Bayes' rule tells us that along with the likelihood we should incorporate the prior, which we assume strongly favors animal, say $p(A) = .999$ and $p(B) = .001$. (Burglars are, thankfully, rare.) For each hypothesis the posterior is proportional to the product of the prior and likelihood, hence

$$p(A|X) \propto p(X|A)p(A) = (.3)(.999) = .2997, \tag{7}$$
$$p(B|X) \propto p(X|B)p(B) = (.8)(.001) = .0008, \tag{8}$$

The denominator in Bayes' rule is the total probability of the data under all hypotheses, here

$$p(X) = p(X|A)p(A) + p(X|B)p(B) = .2997 + .0008 = .3005. \tag{9}$$

Hence the posteriors for animal and burglar are respectively

$$p(A|X) = \frac{p(X|A)p(A)}{p(X)} = \frac{.2997}{.3005} = .9973, \tag{10}$$

$$p(B|X) = \frac{p(X|B)p(B)}{p(X)} = \frac{.0008}{.3005} = .0027, \tag{11}$$

strongly favoring animal. Notice that when comparing the posteriors, we really only need to compare the numerators since the denominators are the same. Hence Bayes' rule is often given in its "proportional" form $p(H|D) \propto p(D|H)p(H)$, in which the denominator is disregarded.

[3] Students are often warned that the likelihood function is *not* a probability distribution, a remark that in my experience tends to cause confusion. In traditional terminology, likelihood is an aspect of the model or hypothesis, not the data, and one refers for example to the likelihood of H (and not the likelihood of the data under H). This is because the term likelihood was introduced by frequentists, who insisted that hypotheses did not have probabilities (see text), and sought a word other than "probability" to express the degree of support given by the data to the hypothesis in question. However, to Bayesians, the distinction is unimportant, since both data and hypotheses can have probabilities, so Bayesians tend (especially recently) to refer to the likelihood of the data under the hypothesis, or the likelihood of the hypothesis, in both cases meaning the probability $p(D|H)$. In this sense, likelihoods are indeed probabilities. However note that the likelihoods of the various hypotheses do not have to sum to one (for example, it is perfectly possible for many hypotheses to likelihood near one given a dataset that they all fit well). In this sense, the sense, the distribution of likelihood over *hypotheses* (models) is certainly not a probability distribution. But the distribution of likelihood over the data for a single fixed model is, in fact, a probability distribution and sums to one.

Parameter estimation

A slightly more complicated application of Bayes' rule involves hypotheses that form a continuous family, that is, where all the hypotheses are of the same general form but differ in the value of one or more continuous parameters. This is often called *parameter estimation* because the observer's goal is to determine, based on the data at hand, the most probable value of the parameter(s), or, more broadly, the distribution of probability of over all possible parameter(s) values (called the *posterior distribution* of the parameter). As a simple example, imagine that we wish to measuring how many milliliters of soda the Dubious Cola company puts in a one-liter bottle. Naturally, there is random variation in every physical process, including both filling the bottles and measuring their contents. So each bottle has a measurement that is the "true" mean μ—the volume the company intends to sell us—plus some random error. Say we measure n bottles, and get a set $X = x_1, x_2 \ldots x_n$ of volumes with a mean $\bar{X}$ of 802 milliliters. What is the true value of μ? Assume that the error around each measurement is normally distributed (a phenomenon so ubiquitous that in the 19th century it was thought of as a natural law, the "Law of Error"; see discussion below). This means that the likelihood, the probability of the observed value conditioned on the value of the parameter, is normal (Gaussian) with standard deviation σ, notated

$$p(x|\mu) = N(\mu, \sigma^2) \tag{12}$$

(In this example for simplicity we'll assume that σ is known and just try to estimate μ.) Because the n measurements are all independent, the entire dataset $X = x_1 \ldots x_n$ has likelihood[4]

$$p(X|\mu) = p(x_1|\mu) \times p(x_2|\mu) \times \ldots \times p(x_{10}|\mu), \tag{13}$$

Classical statistics would say that the best estimate of the "population mean" μ is the value with *maximum likelihood*, which in this case is the sample mean $\bar{X}$, here 802. But Bayes' rule says that in addition to the likelihood, which reflects the information gained from the data, you should incorporate whatever prior information you have about the probable value of the parameter—in this case the assumption that Dubious Cola puts one liter (1000 ml) in a one-liter bottle. Indeed, the optimality of Bayesian inference means that it is in effect *irrational* to ignore this information. In this case it is reasonable to assume that the value of μ is probably about 1000, with (again by assumption) a normal distribution of uncertainty about this value. Narrower distributions would mean stronger biases towards 1000, wider ones weaker biases. (If you really had no idea what value to expect, you could make your prior very wide and flat, in which case it would exert very little influence on the posterior.) Now Bayes' rule tells us that the posterior probability of each value of μ, meaning how believable it is in light of both the data and the prior, is proportional to the product of the prior and likelihood,

$$p(\mu|X) \propto p(X|\mu)p(\mu). \tag{14}$$

[4]In fact, to compute the likelihood of the data we really only need their mean, $\bar{X}$ which is distributed as $N(\mu, \sigma^2/n)$.

This yields a value of $p(\mu|X)$ for every possible value of μ (the posterior distribution) which indicates how strongly we should believe in each value of μ given both the data and our prior beliefs.

Fig. 1 illustrates how the posterior distribution evolves as more data are acquired, and how it relates to the prior and likelihood. The prior is a normal distribution centered at 1000, because that's what we believed a bottle would contain before we started measuring. (In the figure distributions are depicted via their mean plus error bars to indicate one standard deviation; all distributions are normal.) As data is acquired (moving from left to right in the figure), the likelihood is always centered at the sample mean (which is the value that best fits the data so far). But the posterior, which combines the prior with the likelihood via Bayes' rule, is somewhere in between the prior and likelihood—gradually approaching the likelihood, and gradually getting tighter (narrower error bars) as we collect more data and our knowledge gets firmer. That is, the data gradually draw our beliefs away from the prior and towards what the evidence tells us. Thus as we collect more and more data, the posterior distribution increasingly resembles the likelihood distribution. This is often referred to as the likelihood "overwhelming" the prior, and is one of the reasons why in some (though not all) situations the exact choice of prior doesn't matter very much—because as evidence accumulates the prior tends to matter less and less.

The peak of the posterior distribution, the value of the parameter that has the highest posterior probability, is called the *maximum a posteriori* or MAP value. If we need to reduce our posterior beliefs to a single value, this is the most plausible, and casual descriptions of Bayesian inference often imply that Bayes rule dictates that we choose the MAP hypothesis. But remember that Bayes' rule does not actually authorize this reduction; it simply tells how much to believe each hypothesis—that is, the full posterior distribution. In many situations use of the MAP be quite undesirable: for example, broadly distributed posteriors that have many other highly probable values, or multimodal posteriors that have multiple peaks that are almost as plausible as the MAP. Reducing the posterior distribution to a single "winner" discards useful information, and it should be kept in mind that in principle only the entire posterior distribution expresses the totality of our posterior beliefs.

Model selection

Many situations require both discrete hypothesis selection and parameter estimation because the observer has to choose between several qualitatively distinct models, each of which has some number of parameters that must be estimated; this is the problem of *model selection*. Assessing the relative probability of such models can be difficult if, as is often the case, the competing models have different numbers of parameters, because all else being equal models with more parameters have more flexibility to fit the data, since each parameter can act as a "fudge factor" that can improve the fit (increase the likelihood). Classical statistics has very limited tools to deal with this very common situation unless the models are nested (one a subset of the other). But Bayesian techniques can be applied in a straightforward way, the simplest being to consider the ratio of the integrated likelihood of one model relative to that of another, sometimes called the *Bayes factor* (see Kass & Raftery, 1995). This is not the same as comparing the *maximized* likelihood of each model (the likelihood of the model after all its parameters have been set so as to maximize fit to the data). The maximized likelihood ratio, unlike the Bayes factor, considers only the

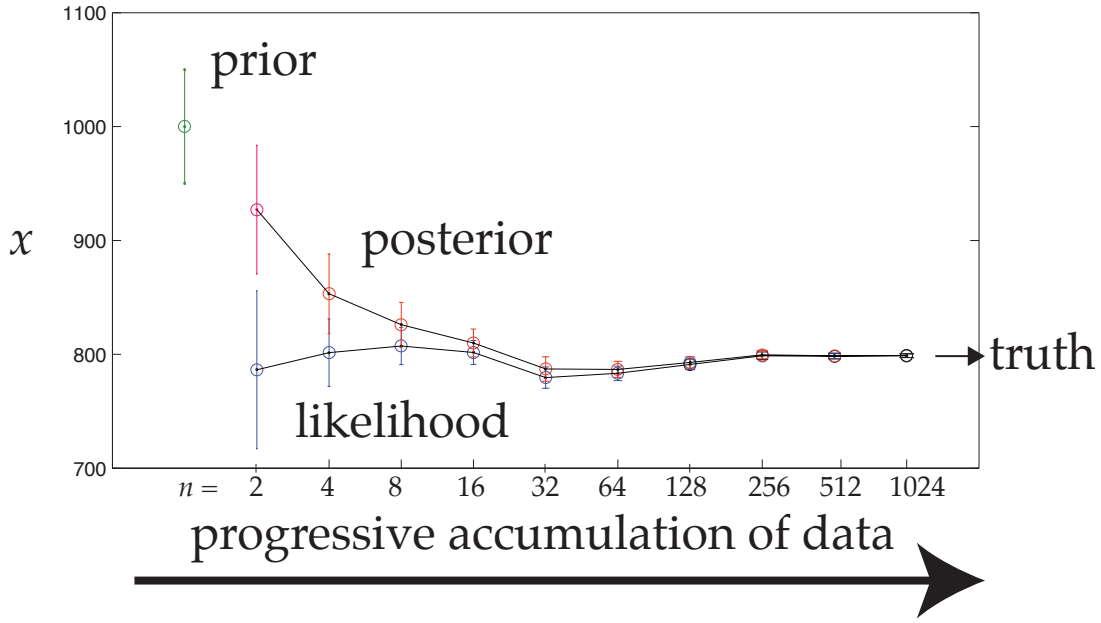

Figure 1. Relationship between prior, likelihood, and posterior distributions as data is accumulated over time. Each distribution here is normal (Gaussian) and is depicted as a point representing the mean, with error bars representing the standard deviation. The observer has a prior centered on $x = 1000$ with a standard deviation of 50. Data are actually generated from a normal centered at $x = 800$. The posterior distribution gradually migrates from the prior, where belief was initially centered, towards the likelihood, where the evidence points. Both likelihood and posterior gradually tighten as more data is acquired.

best fitting parameter settings for each model, which intrinsically favors more complex models (i.e. ones with more parameters) unless a correction is used such as AIC (Akaike, 1974) or BIC (Schwarz, 1978) (see Burnham & Anderson, 2004). But Bayesians argue that no complexity correction is necessary with the use of Bayes factors, because Bayes' rule automatically trades off fit to the data (the likelihood, which tends to benefit from more parameters) with the complexity of the model (which tends to be penalized in the prior; see below). This tradeoff, a version of the *bias-variance* tradeoff that is seen everywhere in statistical inference (see Hastie, Tibshirani, & Friedman, 2001), is quite fundamental to Bayesian inference, because the essence of Bayes' rule is the optimal combination of data fit (reflected in the likelihood) and bias (reflected in the prior).

Computing the posterior

In simple situations, it is sometimes possible to derive explicit formulas for the posterior distribution, as in the examples given above. For example, normal (Gaussian) priors and likelihoods lead to normal posteriors, allowing for easy computation. (Priors and posteriors in the same model family are called *conjugate*.) But in many realistic situations the priors and likelihoods give rise to an unwieldy posterior that cannot be expressed analytically. Then more advanced techniques must be brought to bear, and much of the

modern Bayesian literature is devoted to developing and discovering such techniques. These include Expectation Maximization (EM), Monte Carlo Markov Chains (MCMC), and Bayesian belief networks (Pearl, 1988), each appropriate in somewhat different situations. (See Griffiths and Yuille (2006) for brief introductions to these techniques or Hastie et al., 2001 or Lee, 2004 for more in-depth treatments.) However it should be kept in mind that all these techniques share a common core principle, the determination of the posterior belief based on Bayes' rule.

Bayesian inference in perception

We now turn back to perception, and ask how the Bayesian calculations sketched above can be applied to the fundamental problem of perception, that of estimating the structure of the outside world. The literature on Bayesian perception is now as diverse as it is enormous, and the examples chosen here are intended to be illustrative rather than exhaustive.

Bayesian estimation of surface color

Bayesian inference can be used to estimate perceptual parameters in much the same way it was in the Dubious Cola example. An example is the estimation of color, a classic case of perceptual ambiguity. The reflectance properties of a surface, which determine which wavelengths of light are reflected off the surface in what proportions, are a fixed attribute of the material. But the light that hits our eyes reflects both this attribute, which is what we are trying to determine, and the properties of the light source, which we usually are not. In effect, the quantity of (say) red light that hits our eyes is a product of how much red light is in the light source multiplied by the proportion of red light that the particular surface reflects. Since all we can measure directly is their product, we cannot infer the surface properties—what we care about—without some additional assumptions or tricks. As in all problems of perception, the sensory data is insufficient by itself to disambiguate the properties of the world. The question then is how the brain solves this problem and thus infers the material properties of the surface—thus explaining why red things look red approximately regardless of the color of the light source.

Brainard and Freeman (1997) and Brainard et al. (2006) have proposed a simple Bayesian solution to this problem. First, they assume that the measurement of light amplitude at each frequency is, like the measurement of the volume of Coca-cola, subject to Gaussian error. That is, when our photoreceptors measure the amount of (say) red light reflected off a surface, the measurement is reflects both the true reflectance ρ plus some normally-distributed error. This determines the likelihood function $p(x|\rho)$. But (following Bayes' rule) in order to estimate the true ρ, we need to also consider the prior distribution of $p(\rho)$, that is, the prior probability of that the surface will have the given reflectance ρ prior to considering the image (Fig. 2). Brainard et al. (2006) estimated this by first deriving a low-parameter model of surfaces (that is, finding a small number of parameters that together describe the variation among most surfaces). They then empirically measured the relative frequency of different values of each of these parameters among the surfaces. The results suggest a Gaussian (normal) prior over each of the parameters, meaning that (just as with the volume of Coke bottles) a single mean value with bell-shaped uncertainty

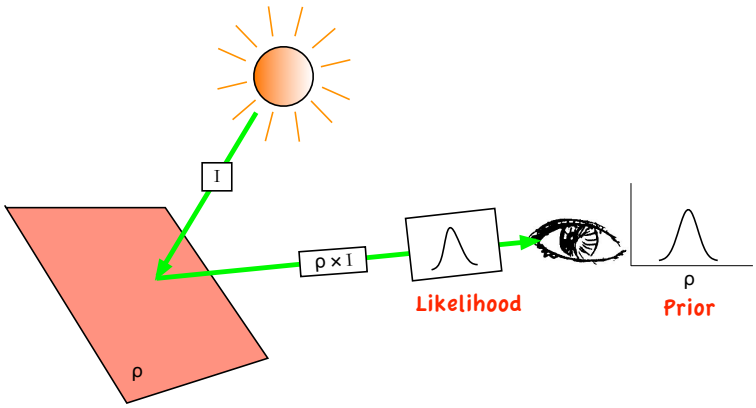

Figure 2. Schematic of the Brainard et al.'s (2006) theory of color estimation. The observer's goal is to infer the true surface reflectance ρ, though the observed light at the given frequency is the product of ρ and the illumination I. The Bayesian solution is to adopt a prior over ρ, and a likelihood function $p(x|\rho)$ that assuming normally distributed noise, which leads to a posterior over potential surface properties.

about the mean. We can then compute the posterior probability of each parameter based on the image data and prior knoweldge about plausible surfaces, to give an estimate of the percevied color of each surface patch. The results show a remarkable agreement with human judgments, suggesting that our color judgments are close to optimal given the uncertainty inherent in the situation.

Bayesian motion estimation

Another basic visual parameter that Bayesian inference can be used to estimate is motion. In everyday vision, we think of motion as a property of coherent objects plainly moving through space, in which case it is hard to appreciate the profound ambiguity involved. But in fact dynamically changing images are generally consistent with many motion interpretations, because the same changes can be interpreted as one visual pattern moving at one velocity (speed and direction), or another pattern moving at another velocity, or many options in between. A simple example of an interpretation failure in this context is the motion of spoked wheels in movies, which (depending on the speed of the wheel relative to the movie frame rate) may sometimes appear to be rotating backwards. The ambiguity is especially pronounced when only a small local region of the image is considered (called the *aperture problem*), and as in many other areas of perception one of the main challenges is the integration of many potentially disparate local estimates of motion into a coherent global estimate.

So the estimation of motion, like that of color, requires deciding which of a range of models is the most plausible interpretation of an ambiguous collection of image data. As such, it can be placed in a Bayesian framework if one can provide (a) a prior over potential motions, indicating which velocities are more a priori plausible and which less, and (b) a likelihood function allowing us to measure the fit between each motion sequence and each potential interpretation. Weiss, Simoncelli, and Adelson (2002) have shown that many phenomena of motion interpretation, including both normal conditions as well as a range

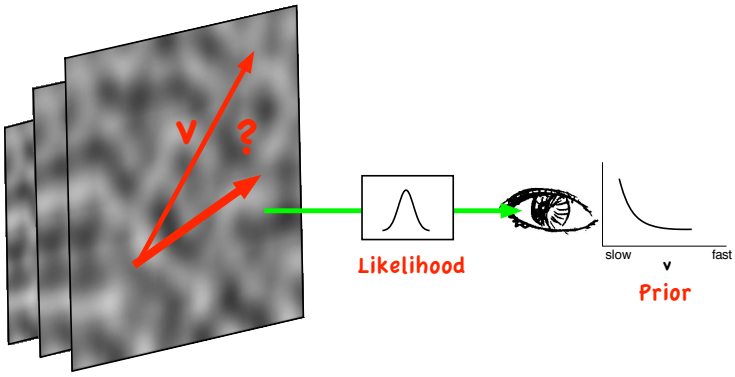

Figure 3. Schematic of Weiss et al.'s (2002) Bayesian model of motion estimation.

of standard motion illusions, are predicted by a simple Baysian model in which (a) the prior favors *slower* speeds over faster ones, and (b) the likelihood is based on conventional Gaussian noise assumptions (Fig. 3). That is, the posterior distribution favors motion speeds and directions that minimize speed while simultaneously maximizing fit to the observed data (leading to the simple slogan "slow and smooth"). The close fit between human percepts and the predictions of the Bayesian model is particularly striking in that in addition to accounting for normal motion percepts, it also systematically explains certain illusions of motions as side-effects of rational inference.

Bayesian contour grouping

The problem of perceptual organization—how to group the visual image into contours, surfaces, and objects—seems at first blush quite different from color or motion estimation, because the property we seek to estimate is not a physical parameter of the world, but a representation of how we choose to *organize* it. Still, Bayesian methods can be applied in a straightforward fashion as along as we assume that each image is potentially subject to many grouping interpretations, but that some are more intrinsically plausible than others (allowing us to define a prior over interpretations), and some fit the observed image better than others (allowing us to define a likelihood function). We can then use Bayes' rule to infer a posterior distribution over grouping interpretations.

A simple example comes from the problem of contour integration, in the question of whether two visual edges belong to the same contour (H_1) or different contours (H_2). Because physical contours can take on a wide variety of geometric forms, practically any observed configuration of two edges is consistent with the hypothesis of a single common contour. But because edges drawn from the same contour tend to be relatively collinear, the angle between two observed edges provides some evidence about how plausible this hypothesis is, relative to the competing hypothesis that the two edge arise from distinct contours. This decision, repeated many times for pairs of edges throughout the image, forms the basis for the extraction of coherent object contours from the visual image.

To formalize this as a Bayesian problem, we need priors $p(H_1)$ and $p(H_2)$ for the two hypotheses, and likelihood functions $p(\alpha|H_1)$ and $p(\alpha|H_2)$ that express the probability of the angle between the two edges (called the *turning angle*) conditioned under each hypothesis.

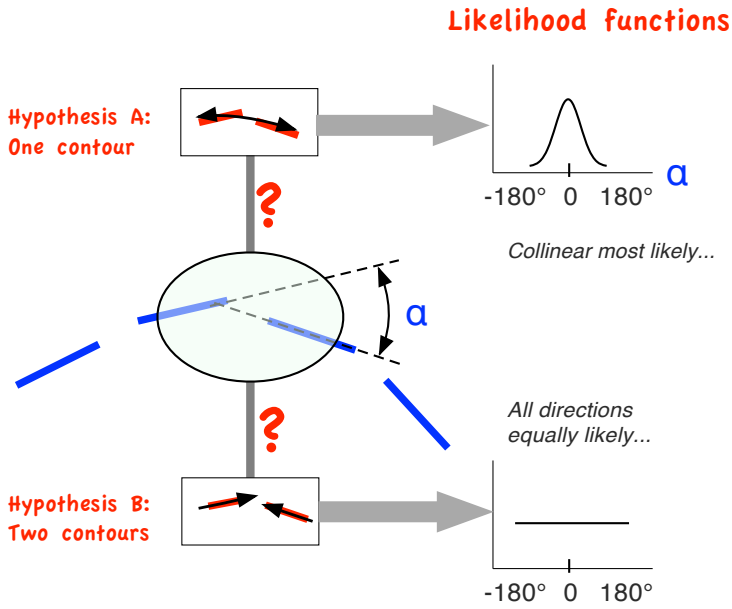

Figure 4. Two edges can be interpreted as part of the same smooth contour (hypothesis A, top) or as two distinct contours (hypothesis B, bottom). Each hypothesis has a likelihood (right) that is a function of the turning angle α; with $p(\alpha|A)$ sharply peaked at 0, but $p(\alpha|B)$ flat.

Several authors have modeled the same-contour likelihood function $p(\alpha|H_1)$ as a normal distribution centered on collinearity (0° turning angle; see Feldman, 1997; Geisler, Perry, Super, & Gallogly, 2001). Fig. 4 illustrates the decision problem in its Bayesian formulation. In essence, each successive pair of contour elements must be classified as either part of the same contour or as parts of distinct contours. The likelihood of each hypothesis is determined by the geometry of the observed configuration, with the normal likelihood function assigning higher likelihood to element pairs that are closer to collinear. The prior (in practice fitted to subjects' responses) tends to favor H_2, presumably because most image edges come from disparate objects. Bayes' rule puts these together to determine the most plausible grouping. Applying this simple formulation more broadly to all the image edge pairs allows the image to be divided up a set of contour elements into a discrete collection of "smooth" contours—that is, contours made up of elements all of which Bayes' rule says belong to the same contour. The resulting parse of the image into contours agrees closely with human judgments (Feldman, 2001). Related models have been applied to contour completion and extrapolation as well (Singh & Fulvio, 2005).

Bayesian perceptual organization

More broadly, perceptual organization in many of its manifestations can be thought of as a Bayesian choice between discrete alternatives. Each qualitatively distinct way of organizing the image constitues an alternative hypothesis. Should a grid of dots be organized into vertical stripes or horizontal ones (Zucker, Stevens, & Sander, 1983)? Should a configuration of dots be grouped into a number of clusters, and if so in what way (Compton & Logan, 1993)? What is the most plausible way to divide a smooth shape into

a set of component parts (Singh & Hoffman, 2001)? Each of these problems can be placed into a Bayesian framework by assigning to each distinct alternative interpretation a prior and a method for determining likelihood.

Each of these problems requires its own unique approach, but broadly speaking a Bayesian framework for any problem in perceptual organization flows from a *generative model* for image configurations (Feldman, Singh, & Froyen, 2012). Perceptual organization is based on the idea that the visual image is generated by regular processes that tend to create visual structures with varying probability, which can be used to define likelihood functions. The challenge of Bayesian perceptual grouping is to discover psychologically reasonable generative models of visual structure.

For example, Feldman and Singh (2006) proposed a Bayesian approach to shape representation based on the idea that shapes are generated from axial structures (skeletons) from which the shape contour is understood to have "grown" laterally. Each skeleton consists of a hierarchically organized collection of axes, and generates a shape via a probabilistic process that defines a probability distribution over shapes (Fig. 5). This allows a prior over skeletons to be defined, along with a likelihood function that determines the probability of any given contour shape conditioned on the skeleton. This in turn allows the visual system to determine the MAP skeleton (the skeleton most likely to have generated the observed shape) or, more broadly, a posterior distribution over skeletons. The estimated skeleton in turn determines the perceived decomposition into parts, with each section of the contour identified with a distinct generating axis perceived as a distinct "part." This shape model is certainly oversimplified relative to the myriad factors that influence real shapes, but the basic framework can be augmented with a more elaborate generative model, and tuned to the properties of natural shapes (Wilder, Feldman, & Singh, 2011). Because the framework is Bayesian, the resulting representation of shape is, in the sense discussed above, optimal given the assumptions specified in the generative model.

Discussion

This section raises several issues that often arise when Bayesian models of cognitive processes are considered.

Simplicity and likelihood from a Bayesian perspective

Bayesian techniques in perception are often associated with what perceptual theorists call the *Likelihood principle*,[5] which is the idea that the brain aims to select the hypothesis that is most likely to be true in the world. Recently Bayesian inference has been held up as the ultimate realization of the principle (Gregory, 2006). Historically, the Likelihood principle has been contrasted with the *Simplicity* or *Minimum Principle*, which holds that the brain will select the simplest hypothesis consistent with sense data (Hochberg & McAlister, 1953;

[5]The Likelihood principle in perception should not be confused with Likelihood principle in statistics, an unrelated idea. The statistical likelihood principle is the idea that the data should influence our belief in a hypothesis only via the probability of that data conditioned on the hypothesis (the likelihood). This principle is universally accepted by Bayesians; indeed the likelihood is the only data-dependent term in Bayes' rule. But it is violated by classical statistics, where, for example, the significance of a finding depends in part on the probability of data that did *not* actually occur in the experiment. (For example, when one integrates the tail of a sampling distribution, one is adding up the probability of many events that did not actually occur.)

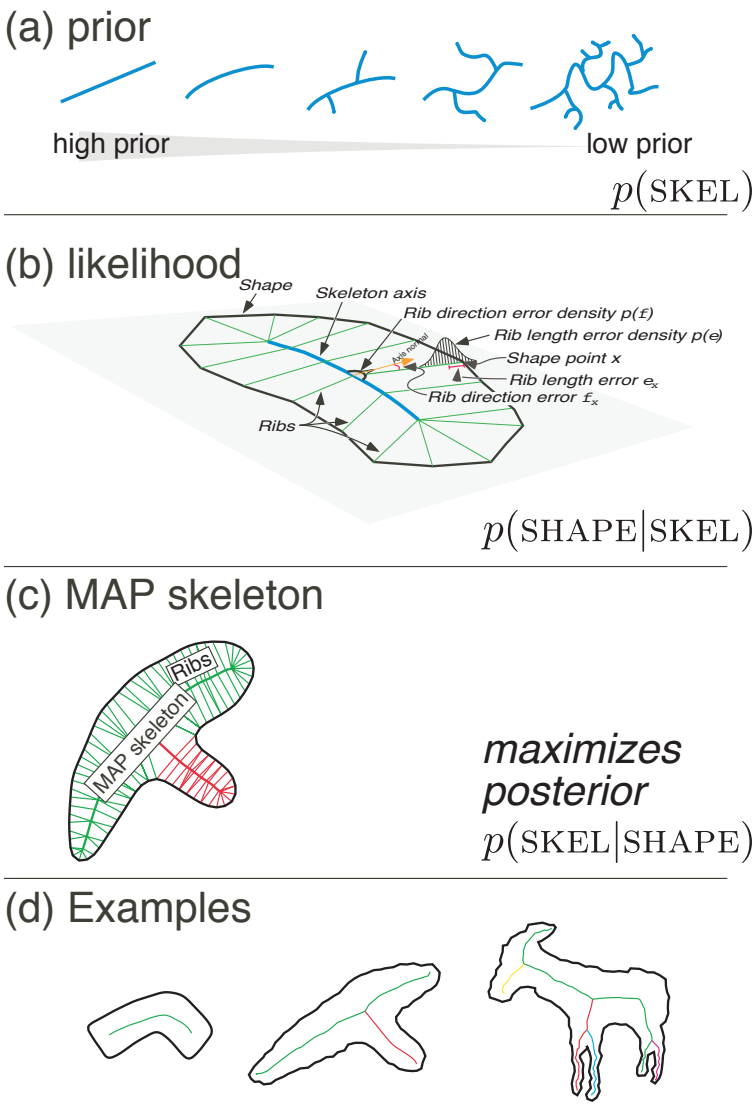

Figure 5. Generative model for shape from Feldman and Singh (2006), giving (a) prior over skeletons (b) likelihood function (c) MAP skeleton, the maximum posterior skeleton for the given shape, and (d) examples of the MAP skeleton.

Leeuwenberg & Boselie, 1988). Simplicity too can be defined in a variety of ways, which has led to an inconclusive debate in which examples purporting to illustrate the preference for simplicity over likelihood, or vice versa, could be dissected without clear resolution (Hatfield & Epstein, 1985; Perkins, 1976).

More recently, Chater (1996) has argued that simplicity and likelihood are two sides of the same coin, for several reasons that stem from Bayesian arguments. First, basic considerations from information theory suggest that more likely propositions are automatically simpler in that they can be expressed in more compact codes. Specifically, Shannon (1948) showed that an optimal code—meaning one that has minimum expected code length—should express each proposition A in a code of length proportional to the negative log probability of A, i.e. $-\log p(A)$. This quantity is often referred to as the *surprisal*, because it quantifies how "surprising" the message is (larger values indicate less probable outcomes), or as the *Description Length* (DL), because it also quantifies how many symbols it occupies in an optimal code (longer codes for more unusual messages). Just as in Morse code (or for that matter approximately in English) more frequently used concepts should be assigned shorter expressions, so that the total length of expressions is minimized on average. Because the proposition with maximum posterior probability (the MAP) also has *minimum* negative log posterior probability, the MAP hypothesis is also the minimum DL (MDL) hypothesis. More specifically, while in Bayesian inference the MAP hypothesis is the one that maximizes the *product* of the prior and the likelihood $p(H)p(D|H)$, in MDL the winning hypothesis is the one that minimizes the *sum* of the DL of the model plus the DL of the data as encoded via the model $(-\log p(H) - \log p(D|H))$, a sum of logs having replaced a product. In this sense the simplest interpretation is necessarily also the most probable—though it must be kept in mind that this easy identification rests on the perhaps tenuous assumption that the underlying coding language is optimal.

More broadly, Bayesian inference tends to favor simple hypotheses even when without any assumptions about the optimality of the coding language.[6] This tendency, sometimes called "Bayes Occam," (after Occam's razor, a traditional term for the preference for simplicity), reflects fundamental considerations about the way prior probability is distributed over hypotheses (see MacKay, 2003). Assuming that the hypotheses H_i are mutually exclusive, then their total prior necessarily equals one ($\sum_i p(H_i) = 1$), meaning simply that the observer believes that one of them must be correct. This in turn means that models with more parameters must distribute the same total prior over a larger set of specific models (combinations of parameter settings) inevitably requiring each model to be assigned a *smaller* prior. That is, more highly parameterized models—models that can express a wider variety of states of nature—necessarily assign lower priors to each individual hypothesis. Hence in this sense Bayesian inference automatically assigns lower priors to more complex models, and higher priors to simple ones, thus enforcing a simplicity metric without any mechanisms designed especially for the purpose.

Though the close relationship between simplicity and Bayesian inference is widely recognized, the exact nature of the relationship is more controversial (see Feldman, 2009 and van der Helm, 2000). Bayesians regard the calculation of the Bayesian posterior as fundamental, and the simplicity principle as merely a heuristic concept whose value

[6] "The simplest law is chosen because it is most likely to give correct predictions" (Jeffreys, 1939/1961), p4).

derives from its correspondence to Bayes' rule. The originators of MDL and information-theoretic statistics (e.g. Akaike, 1974; Rissanen, 1989; Wallace, 2004) take the opposite view, regarding the minimization of complexity (DL or related measures) as the more fundamental principle, and some of the assumptions underlying Bayesian inference as naive (see Burnham & Anderson, 2002; Grünwald, 2005).

Decision making and loss functions

Bayes' rule dictates how belief should be distributed among hypotheses. But a full account of Bayesian decision making requires that we also quantify the *consequences* of each potential decision, usually called the *loss function* (or *utility function* of *payoff matrix*). For example, misclassifying heartburn as a heart attack costs money in wasted medical procedures, but misclassifying a heart attack as heartburn may cost the patient her life. Hence the posterior belief in the two hypotheses (heart attack or heartburn) is not sufficient by itself to make a rational decision: one must also take into account the cost (loss) of each outcome, including both ways of misclassifying the symptoms as well as both ways of classifying them correctly. More broadly, each combination of an action and a state of nature entails a particular cost, usually thought of as being given by the nature of the problem. Bayesian decision theory dictates that the agent select the action that minimizes the (expected) loss—that is, the outcome which (according to the best estimate, the posterior) maximizes the benefit to the agent.

Different loss functions entail different rational choices of action. For example, if all incorrect responses are equally penalized, and correct responses not penalized at all, called *zero-one* loss, then the MAP is the rational choice, because it is the one most likely to avoid the penalty. (This is presumably the basis of the canard that Bayesian theory requires selection of the maximum posterior hypothesis, which is correct only for zero-one loss, and generally incorrect otherwise.) Other loss functions entail other minimum-loss decisions: for example under some circumstances quadratic loss (e.g. loss proportional to squared error) is minimized at the posterior mean (rather than the mode, which is the MAP), and other loss functions entail by the posterior median (Lee, 2004).

Bayesian models of perception have primarily focused on simple estimation without consideration of the loss function, but this is undesirable for several reasons (Maloney, 2002). First, perception in the context of real behavior subserves action, and for this reason in the last few decades the perception literature has evolved towards an increasing tendency to study perception and action in conjunction. Second, more subtly, it is essential to incorporate a loss function in order to understand how experimental data speaks to Bayesian models. Subjects' responses are not, after all, pure expressions of posterior belief, but rather are choices that reflect both belief and the expected consequences of actions. For example, in experiments, subjects implicitly or explicitly develop expectations about the relative cost of right and wrong answers, which help guide their actions. Hence in interpreting response data we need to consider both the subjects' posterior belief and their perceptions of payoff. Most experimental data offered in support of Bayesian models actually shows *probability matching* behavior, that is, responses drawn in proportion to their posterior probability, referred to by Bayesians as *sampling from the posterior.* Again, only zero-one loss would require rational subjects to choose the MAP response on every trial, so probability matching generally rules out zero-one loss (but obviously does not rule out

Bayesian models more generally). The choice of loss functions in real situations probably depend on details of the task, and remains a subject of research.

Loss functions in naturalistic behavioral situations can be arbitrarily complex, and it is not generally understood either how they are apprehended or how human decision making takes them into account. Trommershauser, Maloney, and Landy (2003) explored this problem by imposing a moderately complex loss function on their subjects in a simple motor task; they asked their subjects to touch a target on a screen that was surrounded by several different penalty zones structured so that misses in one direction cost more than misses in the other direction. Their subjects were surprisingly adept at modulating their taps so that expected loss (penalty) was minimized, implying a detailed knowledge of the noise in their own arm motions and a quick apprehension of the geometry of the imposed utility function (see also Trommershauser, Maloney, & Landy, 2008).

Where do the priors come from?

As mentioned above, a great deal of controversy has centered on the origins of prior probabilities. Frequentists long insisted that priors were justified only in presence of "real knowledge" about the relative frequencies of various hypotheses, a requirement that they argued ruled out most uses. A similar attitude is surprisingly common among contemporary Bayesians in cognitive science (see Feldman, in press), many of whom aim to validate priors with respect to tabulations of relative frequency in natural conditions (e.g. Burge, Fowlkes, & Banks, 2010; Geisler et al., 2001). However, as mentioned above, this restriction would limit the application of Bayesian models to hypotheses which (a) can be objectively tabulated and (b) are repeated many times under essentially identical conditions; otherwise objective relative frequencies cannot be defined. Unfortunately, these constraints would rule out many hypotheses which are of central interest in cognitive science, such as interpreting the intended meaning of a sentence (itself a belief, and not subject to objective measurement, and in any event unlikely ever to be repeated) or choosing the "best" way to organize the image (again subjective, and again dependent on possibly unique aspects of the particular image). However, as discussed above, Bayesian inference is not really limited to such situations if (as is traditional for Bayesians) probabilities are treated simply as quantifications of belief. In this view, priors do not represent the relative frequency with which conditions in the world obtain, but rather the observer's *uncertainty* (prior to receiving the data in question) about the hypotheses under consideration.

In Bayesian theory, there are many ways of boiling this uncertainty down to a specific prior. Many descend from the Laplace's *Principle of insufficient reason* (sometimes called the *Principle of indifference*) which holds that a set of hypotheses, none of which one has any reason to favor, should be assigned equal priors. The simplest example of this is the assignment of uniform priors over symmetric options, such as the two sides of a coin or the six sides of a die. More elaborate mathematical arguments can be used to derive specific priors from a generalization of similar symmetry arguments. One is the Jeffrey's prior, which allows more generalized equivalences between interchangeable hypotheses (Jeffreys, 1939/1961). Another is the maximum-entropy prior (Jaynes, 1982), which dictates the use of that prior which introduces the least amount of information—in the technical sense of Shannon—beyond what is known.

Bayesians often favor so-called *uninformative* priors, meaning priors that are as "neu-

tral" as possible; this allows the data (via the likelihood) to be the primary influence on posterior belief. Exactly how to choose an uninformative prior can, however, be problematic. For example, to estimate the success probability of a binomial process, like the probability of heads in a coin toss, it is tempting to adopt a uniform prior over success probability (i.e. equal over the range 0 to 100%).[7] But mathematical arguments suggest that a truly uninformative prior should be relatively peaked at 0 and 100% (the beta(0,0) distribution, sometimes called the Haldane prior; see Lee, 2004). But recall that (as illustrated above), in many situations as data accumulates, the likelihood eventually tends to dominate the posterior. Hence while the source of the prior may be philosophically controversial, in many real situations the actual choice is moot.

More specifically, certain types of simple priors occur over and over again in Bayesian accounts. When a particular parameter x is believed to fall around some value μ, but with some uncertainty that is approximately symmetric about μ, Bayesians routinely assume a Gaussian (normal) prior distribution for μ, i.e. $p(x) \propto N(\mu, \sigma^2)$. Again, this is simply a formal way of expressing what is known about the value of x (that it falls somewhere near μ) in as neutral a manner as possible (technically, this is the maximum entropy prior with mean μ and variance σ^2). Gaussian error is often a reasonable assumption because random variations from independent sources, when summed, tend to yield a normal distribution (the so-called *central limit theorem*).[8] But it should be kept in mind that an assumption of normal error along x does not entail an affirmative assertion that repeated samples of x would be normally distributed—indeed in many situations (such as where x is a fixed quantity of the world, like a physical constant) this interpretation does not even make sense. Such simple assumptions work suprisingly well in practice and are often the basis for robust inference.

Another common assumption is that priors for different parameters that have no obvious relationship are *independent* (that is, knowing the value of one conveys no information about the value of the other). Bayesian models that assume independence among parameters whose relationship is unknown are sometimes called *naive* Bayesian models. Again, an assumption of independence does not reflect an affirmative empirical assertion about the real-world relationship between the parameters, but rather an expression of ignorance about their relationship.

Where do the hypotheses come from?

Another fundamental problem for Bayesian inference is the source of the hypotheses. Bayesian theory provides a method for quantifying belief in each hypothesis, but it does not provide the class of hypotheses themselves, nor any principled way to generate them. Traditional Bayesians are generally content to assume that some member of the hypothesis set lies sufficiently "close" to the truth, meaning that it approximates reality within some acceptable margin of error. However such assumptions are occasionally criticized as naive

[7] Bayes himself suggested this prior, which is now sometimes called Bayes' postulate. But he was apparently uncertain of its validity, and his hesitation may have contributed to his reluctance to publish his Essay, which was published posthumously (see Stigler, 1983).

[8] More technically, the central limit theorem says that the sum of random variables with finite variances tends towards normality in the limit. In practice this means that if x is really the sum of a number of component variables, each of which is random though not necessarily normal itself, then x tends to be normally distributed.

(Burnham & Anderson, 2002).

But the application of Bayesian theory to problems in perception and cognition elevates this issue to a more central epistemological concern. Intuitively, we assume that the real world has a definite state which perception either does or does not reflect. If, however, our hypothesis space does not actually contain the truth—and Bayesian theory provides no reason to believe it does—then it may turn out that *none* of our perceptual beliefs may be literally true, because the true hypothesis was never under consideration (cf. Hoffman, 2009; Hoffman & Singh, in press). In this sense, the perceived world might be *both* a rational belief (in that assignment of posterior belief follows Bayes' rule) and, in a very concrete sense, a grand hallucination (because none of the hypotheses in play are true).

Thus while Bayesian theory provides an optimal method for using all information available to determine belief, it is not magic; the validity of its conclusions is limited by the validity of its premises. This general point is, in fact, well understood by Bayesians, who often argue that all inference is based on assumptions (see Jaynes, 2003; MacKay, 2003). (This is in contrast to frequentists, who aspired to a science of inference free of subjective assumptions.) But it gains special significance in the context of perception, because perceptual beliefs are the very fabric of subjective reality.

Competence vs. performance

Bayesian inference is a rational, idealized mathematical framework for determining perceptual beliefs, based on the sense data presented to the system coupled with whatever prior knowledge the system brings to bear. But it does not, in and of itself, specify computational mechanisms for actually calculating those beliefs. That is, Bayes quantifies exactly how strongly the system should believe each hypothesis, but does not provide any specific mechanisms whereby the system might arrive at those beliefs. In this sense, Bayesian inference is a *competence theory* (Chomsky's term) or a *theory of the computation* (Marr's term), meaning it is an abstract specification of the function to be computed, rather than the means to compute it. Many theorists, concurring with Marr and Chomsky, argue that competence theories play a necessary role in cognitive theory, parallel to but distinct from that of process accounts. Competence theories by their nature abstract away from details of implementation and help connect the computations that experiments uncover with the underlying problem those computations help solve. Conversely, some psychologists denigrate competence theories as abstractions that are irrelevant to real psychological processes (Rumelhart, McClelland, & Hinton, 1986), and indeed Bayesian models have been criticized on these grounds (McClelland et al., 2010; Jones & Love, 2011).

But to those sympathetic to competence accounts, rational models have an appealingly "explanatory" quality precisely because of their optimality. Bayesian inference is, in a well-defined sense, the *best* way to solve whatever decision problem the brain is faced with. Natural selection pushes organisms to adopt the most effective solutions available, so evolution should tend to favor Bayes-optimal solutions whenever possible (see Geisler & Diehl, 2002). For this reason, any phenomenon that can be understood as part of Bayesian model automatically inherits an evolutionary rationale.

Conclusions

In a sense, perception and Bayesian inference are perfectly matched. Perception is the process by which the mind forms beliefs about the outside world on the basis of sense data combined with prior knowledge. Bayesian inference is a system for determining what to believe on the basis of data and prior knowledge. Moreover, the rationality of Bayes means that perceptual beliefs that follow the Bayesian posterior are, in a well-defined sense, optimal given the information available. This optimality has been argued to provide a selective advantage in evolution (Geisler & Diehl, 2002), driving our ancestors towards Bayes-optimal percepts. Moreover optimality helps explain *why* the perceptual system, notwithstanding its many apparent quirks and special rules, works the way it does—because these rules approximate the Bayesian posterior. Moreover, the comprehensive nature of the Bayesian framework allows it to be applied to any problem that can be expressed probabilistically. All these advantages have led to a tremendous increase in interest in Bayesian accounts of perception in the last decade.

Still, a number of reservations and difficulties must be noted. First, to some researchers a commitment to a Bayesian framework seems to involve a dubious assumption that the brain is rational. Many psychologists regard the perceptual system as a hodgepodge of hacks, dictated by accidents of evolutionary history and constrained by the exigencies of neural hardware. While to its advocates the rationality of Bayesian inference is one of its main attractions, to skeptics the hypothesis of rationality inherent in the Bayesian framework seems at best empirically implausible and at worse naive.

Second, more specifically, the essential role of the prior poses a puzzle in the context of perception, where the role of prior knowledge and expectations (traditionally called "top-down" influences) has been debated for decades. Indeed there is a great deal of evidence (see Pylyshyn, 1999) that perception is singularly uninfluenced by certain kinds of knowledge, which at the very least suggests that the Bayesian model must be limited in scope to an encapsulated perception module walled off from information that an all-embracing Bayesian account would deem relevant.

Finally, many researchers wonder if the Bayesian framework is too flexible to be taken seriously, potentially encompassing any conceivable empirical finding. However while Bayesian accounts are indeed quite adaptable, any specific set of assumptions about priors, likelihoods and loss functions provides a wealth of extremely quantitatively specific empirical predictions, which in many specific perceptual domains have been validated experimentally.

Hence notwithstanding all of these concerns, to its proponents Bayesian inference provides something that perceptual theory has never really had before: a "paradigm" in the sense of Kuhn (1962)—that is, an integrated, systematic, and mathematically coherent framework in which to pose basic scientific questions and evaluate potential answers. Whether or not the Bayesian approach turn out to be as comprehensive or empirically successful as its advocates proponents hope, this represents a huge step forward in the study of perception.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Phil. Trans. of the Royal Soc. of London*, *53*, 370–418.

Brainard, D. H., & Freeman, W. T. (1997). Bayesian color constancy. *Journal of the Optical Society of America A*, *14*, 1393–1411.

Brainard, D. H., Longere, P., Delahunt, P. B., Freeman, W. T., Kraft, J. M., & Xiao, B. (2006). Bayesian model of human color constancy. *J Vis*, *6*(11), 1267–1281.

Burge, J., Fowlkes, C. C., & Banks, M. S. (2010). Natural-scene statistics predict how the figure-ground cue of convexity affects human depth perception. *J. Neurosci.*, *30*(21), 7269–7280.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multi-model inference: A practical information-theoretic approach*. New York: Springer.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference : Understanding AIC and BIC in model selection. *Sociological Methods & Research 2004 33: 261*, *33*(2), 261–304.

Chater, N. (1996). Reconciling simplicity and likelihood principles in perceptual organization. *Psychological Review*, *103*(3), 566–581.

Compton, B. J., & Logan, G. D. (1993). Evaluating a computational model of pereptual grouping by proximity. *Perception & Psychophysics*, *53*(4), 403–421.

Cox, R. T. (1961). *The algebra of probable inference*. London: Oxford Universitiy Press.

de Finetti, B. (1970/1974). *Theory of probability*. Torino: Giulio Einaudi. (Translation 1990 by A. Machi and A. Smith, John Wiley and Sons)

Earman, J. (1992). *Bayes or bust? : a critical examination of bayesian confirmation theory*. MIT Press.

Feldman, J. (1997). Curvilinearity, covariance, and regularity in perceptual groups. *Vision Research*, *37*(20), 2835–2848.

Feldman, J. (2001). Bayesian contour integration. *Perception & Psychophysics*, *63*(7), 1171–1182.

Feldman, J. (2009). Bayes and the simplicity principle in perception. *Psychological Review*, *116*(4), 875–887.

Feldman, J. (in press). Tuning your priors to the world. *Topics in Cognitive Science*.

Feldman, J., & Singh, M. (2006). Bayesian estimation of the shape skeleton. *Proceedings of the National Academy of Science*, *103*(47), 18014–18019.

Feldman, J., Singh, M., & Froyen, V. (2012). *Perceptual grouping as Bayesian mixture estimation.* (Forthcoming.)

Geisler, W. S., & Diehl, R. L. (2002). Bayesian natural selection and the evolution of perceptual systems. *Philosophical Transactions of the Royal Society of London B*, *357*, 419–448.

Geisler, W. S., Perry, J. S., Super, B. J., & Gallogly, D. P. (2001). Edge co-occurrence in natural images predicts contour grouping performance. *Vision Research*, *41*, 711–724.

Gregory, R. (2006). Editorial essay. *Perception*, *35*, 143–144.

Griffiths, T. L., & Yuille, A. L. (2006). A primer on probabilistic inference. *Trends in Cognitive Sciences*, *10*(7).

Grünwald, P. D. (2005). A tutorial introduction to the minimum description length principle. In P. D. Grünwald, I. J. Myung, & M. Pitt (Eds.), *Advances in minimum description length: Theory and applications.* Cambridge, MA: MIT press.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer.

Hatfield, G., & Epstein, W. (1985). The status of the minimum principle in the theoretical analysis of visual perception. *Psychological Bulletin*, *97*(2), 155–186.

Hochberg, J., & McAlister, E. (1953). A quantitative approach to figural "goodness". *Journal of Experimental Psychology*, *46*, 361–364.

Hoffman, D. D. (2009). The user-interface theory of perception: Natural selection drives true

perception to swift extinction. In S. Dickinson, M. Tarr, A. Leonardis, & B. Schiele (Eds.), *Object categorization: Computer and human vision perspectives.* Cambridge: Cambridge University Press.

Hoffman, D. D., & Singh, M. (in press). Computational evolutionary perception. *Perception*.

Howie, D. (2004). *Interpreting probability: Controversies and developments in the early twentieth century*. Cambridge: Cambridge University Press.

Jaynes, E. T. (1982). On the rationale of maximum-entropy methods. *Proceedings of the I.E.E.E.*, *70*(9), 939–952.

Jaynes, E. T. (2003). *Probability theory: the logic of science*. Cambridge: Cambridge University Press.

Jeffreys, H. (1939/1961). *Theory of probability (third edition)*. Oxford: Clarendon Press.

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

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *9*, 773–795.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology*, *55*, 271—304.

Knill, D. C., Kersten, D., & Yuille, A. (1996). Introduction: a Bayesian formulation of visual perception. In D. C. Knill & W. Richards (Eds.), *Perception as Bayesian inference* (pp. 123–162). Cambridge: Cambridge University Press.

Knill, D. C., & Richards, W. (Eds.). (1996). *Perception as Bayesian inference*. Cambridge: Cambridge University Press.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. U. Chicago Press.

Lee, P. (2004). *Bayesian statistics: an introduction (3rd ed.)*. Wiley.

Leeuwenberg, E. L. J., & Boselie, F. (1988). Against the likelihood principle in visual form perception. *Psychological Review*, *95*, 485–491.

MacKay, D. J. C. (2003). *Information theory, inference, and learning algorithms*. Cambridge: Cambridge University Press.

Maloney, L. T. (2002). Statistical decision theory and biological vision. In D. Heyer & R. Mausfeld (Eds.), *Perception and the physical world: Psychological and philosophical issues in perception* (pp. 145–189). New York: Wiley.

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T., Seidenberg, M. S., et al. (2010). Letting structure emerge: Connectionist and dynamical systems approaches to understanding cognition. *Trends Cogn. Sci.*, *14*, 348–356.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. San Mateo, CA: Morgan Kauffman.

Perkins, D. (1976). How good a bet is good form? *Perception*, *5*, 393–406.

Pylyshyn, Z. (1999). Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behav Brain Sci*, *22*(3), 341–365.

Ramachandran, V. S. (1985). The neurobiology of perception. *Perception*, *14*, 97–103.

Rissanen, J. (1989). *Stochastic complexity in statistical inquiry*. Singapore: World Scientific.

Rumelhart, D. E., McClelland, J. L., & Hinton, G. E. (1986). *Parallel distributed processing: explorations in the microstructure of cognition*. Cambridge, Massachusetts: MIT Press.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics 6*, *2*, 461–464.

Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*, 379–423.

Singh, M., & Fulvio, J. M. (2005). Visual extrapolation of contour geometry. *PNAS*, *102*(3), 939–944.

Singh, M., & Hoffman, D. D. (2001). Part-based representations of visual shape and implications for visual cognition. In T. Shipley & P. Kellman (Eds.), *From fragments to objects: segmentation and grouping in vision, advances in psychology, vol. 130* (pp. 401–459). New York: Elsevier.

Stigler, S. M. (1983). Who discovered Bayes's theorem? *The American Statistician*, *37*(4), 290–296.

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Harvard University Press.

Trommershauser, J., Maloney, L. T., & Landy, M. S. (2003). Statistical decision theory and the selection of rapid, goal-directed movements. *J Opt Soc Am A Opt Image Sci Vis*, *20*(7), 1419–1433.

Trommershauser, J., Maloney, L. T., & Landy, M. S. (2008). Decision making, movement planning and statistical decision theory. *Trends Cogn. Sci.*, *12*(8), 291–297.

van der Helm, P. (2000). Simplicity versus likelihood in visual perception: From surprisals to precisals. *Psychological Bulletin*, *126*(5), 770–800.

Wallace, C. S. (2004). *Statistical and inductive inference by minimum message length*. Springer.

Weiss, Y., Simoncelli, E. P., & Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nat. Neurosci.*, *5*(6), 598–604.

Wilder, J., Feldman, J., & Singh, M. (2011). Superordinate shape classification using natural shape statistics. *Cognition*, *119*, 325–340.

Yuille, A. L., & Bülthoff, H. H. (1996). Bayesian decision theory and psychophysics. In D. C. Knill & W. Richards (Eds.), *Perception as Bayesian inference* (pp. 123–162). Cambridge: Cambridge University Press.

Zucker, S. W., Stevens, K. A., & Sander, P. (1983). The relation between proximity and brightness similarity in dot patterns. *Perception and Psychophysics*, *34*(6), 513–522.