Bayesian Modeling of Sensory Cue Combinations

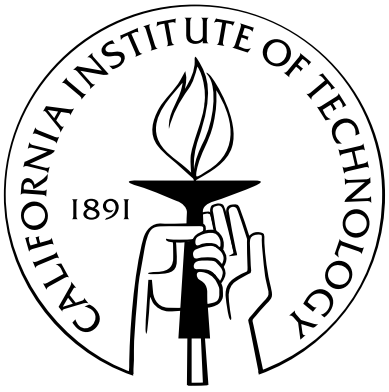
Thesis by

Ulrik Ravnsborg Beierholm

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy



California Institute of Technology

Pasadena, California

2007

(Defended May 15, 2007)

Acknowledgements

> But when the Rabbit actually took a watch out of its waistcoat-pocket and looked at it and then hurried on, Alice started to her feet, for it flashed across her mind that she had never before seen a rabbit with either a waistcoat-pocket, or a watch to take out of it, and, burning with curiosity, she ran across the field after it and was just in time to see it pop down a large rabbit-hole, under the hedge. In another moment, down went Alice after it!
>
> –*Alice's Adventure in Wonderland*, Lewis Carroll, 1916.

Science can be quite the adventure, and I have a long list of people whose help along the way I am grateful for. My advisor, Steven Quartz, for giving me the opportunity to work on what I considered interesting. Each member of my committee have helped in their own way. Ladan Shams has been a driving force in collaborating on most of my experiments and constantly pushing new ideas on me. Peter Bossaerts has helped with some of the theoretical ideas in this thesis and has been the main force behind the Bayesian learning studies. John O'Doherty has given advice on the fMRI studies that we are still working on, and Erik Winfree has encoraged me through many discussions of neural computation.

My other collaborators, Wei Ji Ma and Konrad Koerding have helped to make these projects a true pleasure to work on. I also wish to thank Rajesh Rao at University of Washington for originally suggesting I work on this problem.

The Quartz Lab members have been extremely helpful in many discussions, especially Kerstin Preuschoff, Antoine Bruguier and Cedric Anen. Likewise have the other CNS students from the 2001 incoming year by their suggestions and encouragement.

Lastly I especially want to thank my family, my parents Niels and Rita, my brother Anders and my wife Amy. I am eternally grateful for their support.

Financial support for this work was provided by the David and Lucile Packard Foundation and the Gordon and Betty Moore Foundation.

Abstract

We are constantly bombarded by sensory input, from multiple sources. How does the human brain process all this information? How does it decide what information to combine and what to keep segregated? If a visual and auditory stimulus originates from the same cause, say a baseball hitting a bat, there are advantages to combining the information to a single percept.

The Bayesian framework is an obvious way to approach this problem. Given sensory input, the optimal observer would try to infer what sources created the stimuli.

We developed a model based on Bayesian principles that makes no assumptions about the type of correlations between sources in the world, and found it in excellent correspondence with experiments on human subjects.

Further, it is possible to place this problem in the more specific frame of Causal Inference, which has previously only been applied to cognitive problems. Causal inference tries to infer the hidden causal structure from the observables, implying that we constantly are inferring the causal structure of the world surrounding us. This can be thought of as a more constrained approach than our previous model with fewer parameters.

We performed a number of psychophysics experiments and found the subjects' responses to be in excellent agreement with the model's predictions. Further predictions from the model, such as independence of the Likelihoods and Priors, were tested and confirmed by experimental data.

These studies show that human perception is in accordance with Bayesian inference and implies a commonality between perceptual and cognitive processing.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

There is a tradition going back at least to the nineteen forties [McCulloch and Pitts, 1943] of thinking of the brain as a computational device, a number cruncher that when fed some string of numbers would output some other numbers in the form of words, movements, etc. This tradition goes hand in hand with our understanding of Boolean logic and the computers that arose from it. Given a statement (S) as input we do a logical transformation to produce an output (NOT S). The logical transformation is clear, unambiguous, and - most importantly - deterministic. Which is fine when your input and logical circuits are clear, unambiguous and deterministic.

Not so for the nervous system. The input is generally noisy, ambiguous, and stochastic, and the circuits performing the calculations are unreliable and stochastic themselves [Kandel et al., 2000]. Imagine designing a computer whose components would be dependent on temperature and humidity, whose energy source would be varying in frequency and voltage, its input varying over time and bound to give slightly different answers when queried. If you think your Windows computer is unstable now, imagine what it would be like running on this system!

Alternative solutions are obviously needed, and one of the ideas proposed elsewhere [von Helmholtz, 1865-1867] and supported in this work is the idea of using probabilistic inference. Instead of reacting to the input received, the nervous system would try to infer the most likely cause of the input and base its action upon this. It is a radically different way of thinking as it implies that the brain is constantly internally recreating the world based partly on its noisy input, partly on what it expects to experience.

In fact what this thesis hopes to convince the reader of, is that many of the illusions that we experience can in fact be accredited to the brain having an internal prior model of

the world, which constantly influences the perceived surroundings. In most situations such a model is helpful in recreating the world realistically, but in some specific artificial setups, this prior influence can be detrimental to the task.

More specifically we will here focus on Bayesian modeling of audio-visual percepts for a large range of congruent and incongruent stimuli. Chapter 2 will provide the theoretical background for Bayesian modeling and Chapter 3 will give an example of how to apply the theory to one audio-visual illusion. In Chapter 4 we will start debating a specific type of Bayesian inference, causal inference, and in Chapter 5 show how to apply this to perceptual results. Chapter 6 and 7 continue using this model for further predictions, and Chapter 8 wraps everything up by comparing the restricted causal inference model with a less restricted Bayes' model. Finally the Appendix shows two examples of future work that can be done in this field and includes some mathematics and details that there was not room for in the main text.

Chapters 3, 5, 6, and 7 are each based on papers that have either been published, submitted or are currently under revision. Experiments in Chapter 3 were performed by Ladan Shams, the analysis in Section 5.4 and Figures 5.1 and 5.4 were by Konrad Koerding, the work in Appendix B was done together with Cedric Anen.

Chapter 2

The Optimal Decision

Given uncertain information, how do we make decisions? Are there general rules we can utilize? The theory of Bayesian inference is presented and applied to several graphical models. The application of these ideas to the nervous system is also discussed.

2.1 Noisy environment

We are always forced to make decisions based on uncertain information. The only things we can know with certainty are axiomatic; logic and mathematics. Any statement about the state of the world will naturally involve some amount of uncertainty.

For example, our current measurement of time is given by the cesium atomic clock. SI defines the second as 9,192,631,770 cycles of the radiation released by the transition between two electron spin energy levels of the ground state of the Cs-133 atom. However in recognition of the uncertainty of the measurement of even this, the International Atomic Time is calculated by averaging over measurements from roughly 300 laboratories around the world (http://www.npl.co.uk/time/).

Our everyday world we may perceive as certain under regular everyday activity, but it does not take large changes in our environment, say taking a walk at dusk, for us to realize that our sensory estimates are uncertain. Reaching for an object always requires estimating the distance to the object, an estimate that can occasionally be off considerably.[1]

How do we optimally make decisions given this uncertainty? Do there exist general rules for information processing that the nervous system may or may not have discovered how to use? In the following we will discuss general rules for calculating such a statistically

[1] Just try reaching for that cup of coffee/tea next to you with your eyes closed. Come on, I dare you.

optimal approach, based on simple assumptions.

2.1.1 Probabilistic description

The first step is to realize that if we do not have certain information, we need to use probabilistic estimates. Given a property A we will talk about $P(A)$ as the probability of A taking a specific value. A can be binary value (e.g., rain or no rain), a continuous value (amount of rain in an hour), or a discrete value (number of cars crashing in the rain on the 110 freeway).[2] In the following we will refer to A as a discrete variable, with the knowledge that we can generally substitute a continuous value by, e.g., exchanging integration for summation.

2.1.2 Combining information

Imagine that you are receiving two estimates of the same variable x. Which one should you believe? If you know that one of the estimates is completely wrong, you can choose to completely ignore it and only rely on the other one. More generally speaking you can combine the information, by assigning weights to each contribution:

$$\hat{x} = w_1\hat{x_1} + w_2\hat{x_2}. \tag{2.1}$$

Determining the weights, w_1, w_2, is now the tricky part. You need to have a predefined notion of how reliable each cue is or be able to estimate it by some other means.

2.2 Bayesian inference

If we are willing to venture into the realm of probabilistic solutions there is a simple, and in fact optimal solution, given by Bayes' theorem. Denoting two quantities A and B, their joint probability, that is the probability of them both taking specific values can be split as follows:

$$P(A, B) = P(B|A)P(A) = P(A|B)P(B) \tag{2.2}$$

[2]For the continuous case we are dealing with probability densities, as asking about the probability of it raining exactly 3.14159 mm in a given day, does not make as much sense as asking if it rains more than 3 mm and less than 4 mm.

from which we derive Bayes' theorem [Berger, 1980]

$$P(A|B) = P(B|A)P(A)/P(B). \tag{2.3}$$

This seemingly simple expression allows us to make some strong statements. Given our knowledge of quantity B, we can learn about the hidden quantity A from our knowledge of B's dependence on A and the unconditional knowledge of A.

$P(A)$ is usually referred to as the prior, as it is unconditional on the observed estimate B, and is therefore prior to any observations. $P(A|B)$ is referred to as the posterior probability, since it reflects the new estimate, given the newly available information.

As Bayes' theorem is an inference rule derived directly from the laws of probability with a minimum number of assumptions[3] , it is statistically optimal; any other way of combining information (e.g., adding distributions) can either be shown to be statistically sub-optimal, or equivalent to Bayes' theorem.

2.3 Examples

A few examples are appropriate here. Imagine that I am sitting in my office in the basement, wondering whether or not I need to water the plants at home. I have no windows and the only website I have access to is the Highway Patrol's statistics of number of car crashes on the freeway for different weather conditions. The website also tells me that there were 85 crashes today. Given the number of crashes, what is the probability that it rained that day?

According to Bayes' theorem, this will be the probability of the observed number of car crashes on an average rainy day, times the average probability of rain in my town, divided by the probability of that number of crashes. This looks simpler in equation form:

$P(R|C = 85) = \frac{P(C{=}85|R)P(R)}{P(C{=}85)}$.

The number of crashes on the freeway is here a discrete number while Rain/No rain is clearly binary. By looking up the numbers in the highway patrols records, I should be able to estimate the probability of it having rained today. Whether I happen to live in Los Angeles or Seattle can have quite an influence on my estimate, as $P(R)$ can vary.

[3]Obviously we here assume $P(B)$ is non-zero, but as our posterior is conditional upon it, $P(B) = 0$ would create a logical inconsistency.

A more typical example comes from the pharmaceutical industry, where the prior can also have a large influence. Imagine a disease D that affects 1 percent of the general population $P(D+) = 0.01$. A new diagnosis is developed that is able to detect an early stage of the disease with 99 percent correct positives $P(M+|D+) = 0.99$ and only 1 percent false positives $P(M+|D-) = 0.01$. Although this seems good, if we do the math according to Bayes' theorem we find that the probability of having the disease given the positive result of the test is

$P(D+|M+) = \frac{P(M+|D+)P(D+)}{P(M+|D+)P(D+)+P(M+|D-)P(D-)} = \frac{0.99*0.01}{0.99*0.01+0.01*0.99} = 0.5.$

In other words, even if the test is positive you still only have a 50 percent chance of having the disease. Obviously such a measurement should be followed up by more testing of the subject. This example clearly shows the importance of taking the prior probability into account when performing such inference.

Bayes' theorem is also able to explain counterintuitive results such as the Monty Hall paradox [4].

2.4 Utility function

The last step in any Bayesian decision is applying the utility function and performing the action. Knowing the probability distribution of a quantity X given V, $P(X|V)$, does not tell you how to react towards it. In a simple case where your task is to point to your perceived location $\hat{x}$, you need to know how to estimate $\hat{x}$ from your posterior, dependent on what the cost/task is of being wrong [Berger, 1980, Trommershauser et al., 2003].

$$\hat{x} = \int U(X)P(X|V)dX. \tag{2.4}$$

A specific cost function specifies a specific strategy, e.g., for an error function that only provides utility for the correct answer, $U(x) = \delta(x - x_{real})$ the optimal strategy is to find the maximum of the posterior (Maximum Posterior or MAP). Another example is for an error function that rewards getting as close to the correct answer as possible, $U(x) = (x - x_{real})^2$, in which case the optimal strategy is to pick the mean of the posterior. Of course for Gaussian distributions, the mean is at the maximum, but as we will see later we don't always get normally distributed posteriors.

[4] http://mathworld.wolfram.com/MontyHallProblem.html

A very different alternative is to sample the posterior, i.e., responding randomly with the probability of choosing a response given by its posterior probability, i.e., probability matching [Fantino, 1998]. This type of response is not optimal, but has been used previously to explain variability in subjects' responses [Mamassian and Landy, 2001] and there is evidence that subjects use the rule under certain conditions [Shanks et al., 2002].

In the following, unless otherwise noted, we will assume a utility function that rewards getting as close as possible to the correct answer and which therefore yields the mean of the posterior as the optimal decision.

2.5 Maximum Likelihood and the role of the prior

The use of Bayesian inference has only become common during the last 50 years despite Reverend Thomas Bayes describing it almost 250 years ago [Bayes, 1764]. It could be argued that the reason for this is that traditional estimation of parameters in statistics can be considered a special case of Bayes' theorem. Indeed if in equation (2.3) we assume a uniform prior, $P(A)$, we find that Bayes' theorem turns into

$$P(A|B) \sim P(B|A) \tag{2.5}$$

and if we take the maximum of the posterior, we are instead taking the maximum of the likelihood. Bayes' theorem with a flat prior is therefore mathematically equivalent to the Maximum Likelihood (ML) method, that does not include a prior. The ML is a very simple and intuitive approach and although only an approximation in the Bayesian scheme, therefore highly used. However, ML does not have a way of combining several estimates, nor does it allow you to choose your utility function and decision rule .

Occasionally, any model with a flat prior will be referred to as a maximum likelihood model, regardless, e.g., of the utility function used, we will reserve the term for the model mentioned above. For a more detailed discussion of the differences between maximum likelihood and Bayesian inference, see Berger [1980].

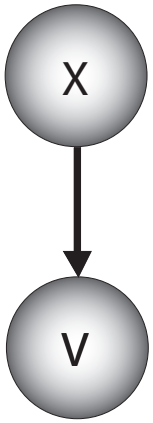


Figure 2.1: A graphical representation of the model in equation (2.6). The generative model assumes that the source X generates signal V, which we measure.

2.6 Optimal inference

2.6.1 Optimal inference, 1 cue

We will now examine the situation where we wish to know about a single quantity X given some visual information V. We can here use Bayes' theorem from above:

$$P(X|V) = P(V|X)P(X)/P(V) \tag{2.6}$$

We will assume that our quantities $P(X|V)$ and $P(X)$ are normally distributed, i.e., that their probabilities are given by Gaussian distributions, with mean μ_V and μ_X and variances σ_V^2 and σ_X^2. The validity of this assumption is obviously dependent on the particulars of the system, but is - due to the central limit theorem - often very reasonable. Making this assumption allows us to take advantage of a wonderful property of normal distributions; their product is also a normal distribution with new variance σ_N^2 and mean μ_N given by

$$\sigma_N^2 = \left(\frac{1}{\sigma_V^2} + \frac{1}{\sigma_X^2}\right)^{-1} \tag{2.7}$$

$$\mu_N = \frac{\mu_V/\sigma_V^2 + \mu_X/\sigma_X^2}{1/\sigma_V^2 + 1/\sigma_X^2} = \sigma_N^2\left(\frac{\mu_V}{\sigma_V^2} + \frac{\mu_X}{\sigma_X^2}\right). \tag{2.8}$$

As discussed, the mean of the posterior is also the optimal decision given a utility function that tries to minimize the squared deviation from the correct answer. We now notice that the expression for the mean has the exact form of the weighted average that we discussed earlier

$$\hat{x} = w_1\hat{x_1} + w_2\hat{x_2} \tag{2.9}$$

with weights given by

$$w_1 = \frac{1/\sigma_V^2}{1/\sigma_V^2 + 1/\sigma_X^2} \tag{2.10}$$

and

$$w_2 = \frac{1/\sigma_X^2}{1/\sigma_V^2 + 1/\sigma_X^2}. \tag{2.11}$$

We therefore have our answer, Bayes' theorem optimally provides the weights for the combination of your two sources of information, in this case a conditional (likelihood) and unconditional (prior) source of information.

We shall see how this extends to more complex situations with several cues.

2.6.2 Optimal inference, 2 cues

Imagine now that you have two cues of conditional information about the same source, say auditory A and visual V. You wish to estimate what the probability is of the source taking a certain value, given these pieces of information. We can extend the scheme from above into a new graphical scheme (see Figure 2.2).

$$P(X|A,V) = \frac{P(A,V|X)P(X)}{P(A,V)} = \frac{P(A|X)P(V|X)P(X)}{P(A,V)} \tag{2.12}$$

We here take advantage of the fact that A and V are conditionally independent, given X, $P(A,V|X) = P(A|X)P(V|X)$. The posterior is now given by the normalized product of two likelihood functions and a prior. The extra piece of information about the source X, given by A, can therefore directly be included by simply multiplying the likelihood together with the previous result. In fact, we could continue adding further cues this way, by simply multiplying all the likelihoods together with the prior. But what if we care about more than just one quantity, X?

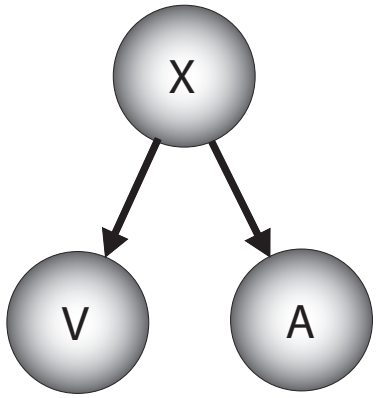


Figure 2.2: Graphical representation of the model in equation (2.12). The source X causes two signals, A and V, which we have access to.

2.6.3 Multiple causes

In real life we are exposed to a multitude of stimuli, that we do not want to all combine. The brain needs to be able to separate estimates when they are from independent sources and integrate them when they are information about the same source. One possible model that enables this, assumes an unspecified interaction between the sources in the world. The interaction can be strong, as in a single source, or it can be weak to the point that there is no correlation between the sources. In the following we will assume that we can encapsulate this interaction in the prior probability $P(X_A, X_V)$

$$P(X_A, X_V|A, V) = \frac{P(A|X_A)P(V|X_V)P(X_A, X_V)}{P(A, V)}. \tag{2.13}$$

Note that we make the same assumption of independence of likelihoods as above.

In fact the two previous models can be considered special cases of this model, with $P(X_A, X_V)$ taking specific shapes. For the single cue, single source case, we get this if we assume a decomposable prior $P(X_A, X_V) = P(X_A)P(X_V)$, implying no interaction between the sources. In this case we have two separate situations, that can be treated separately (no arrow between X_A and X_V in Figure 2.3).

For the case with one cause but two cues, if we set $X_A = X_V$, we find the expression in Section 2.2. That is, the prior $P(X_A, X_V)$ is purely diagonal.

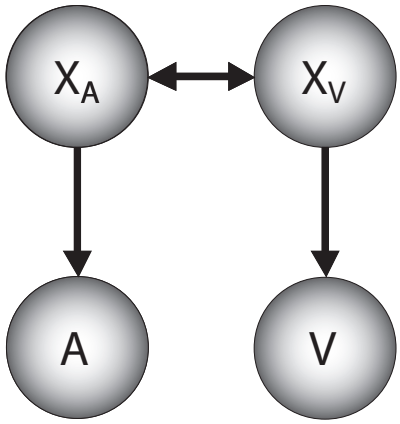


Figure 2.3: A very general model that assumes interaction between two sources X_A and X_V, each giving rise to a signal A and V

The current model is therefore a super-set of the previous models, with more parameters and potentially more power. For our purposes this is the most encompassing model and we will later examine restrictions on it.

2.7 Optimal decisions in the nervous system

During the last 15 years several behavioral studies have looked at to what degree the nervous system uses such optimal techniques [Rao et al., 2002]. In one study Jacobs et al. [Jacobs, 1999] examined how humans estimate the depth of an object. They presented subjects with an image of a elliptical cylinder and asked subjects to estimate the depth of it, using two cues; a texture and a motion cue. For the texture cue, isotropic circles would be projected unto the cylinder and would thus be deformed from the subjects viewpoint dependent on the depth of the cylinder (see Figure 2.4). The motion cue was created by moving dots of light across the horizontal, their motion again being deformed by the depth of the cylinder.

By measuring the subjects' performance on each cue independently, the researchers could predict how subjects would combine the two cues according to equation (2.9). They found that the subjects combination of the two cues was very close to optimal, however

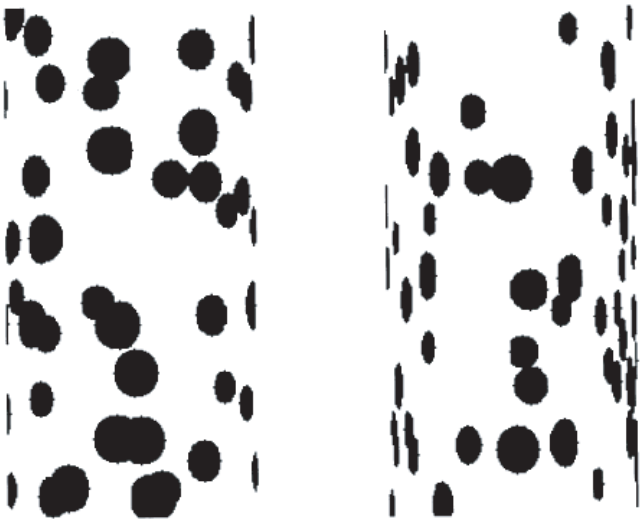


Figure 2.4: Two examples of the texture cue from [Jacobs and Fine, 1999]. The column on the left has a depth equal to its width, the column on the right is deeper than its width.

they never tested any influence of priors.

In another study Weiss et al. tested human perception of visual motion. Under low-contrast conditions they presented moving stimuli [Weiss et al., 2002] and asked subjects to compare the speed of the object to a standard stimulus in a 2-alternative forced-choice task (2AFC [Macmillan and Crelman, 1991]). As they varied the contrast of the stimulus subjects indicated perceiving slower movement, though the speed was held constant. The experimenters chose to interpret this in terms of a 1-cue model (similar to Figure 2.1), where subjects have a larger prior probability for lower speeds. As the contrast was decreased, the reliability of the visual cue would therefore also decrease, causing subjects to rely more on the prior. In a later study they compared subjects performance directly to a Bayesian model and found it in good correspondence [Stocker and Simoncelli, 2006].

However no studies so far have considered the much more general case of cue interaction as in Figure 2.3; cues to separate sources can be interacting in ways different than full fusion, yet not completely independent.

2.8 Summary

As we have seen here, there are ways to perform optimally given uncertainty in the environment. A simple rule, Bayes' theorem allows us, given a structure and and prior, to use the received information to make optimal inferences and decisions. A few studies have examined to what degree the human nervous system utilizes this information optimally, for the case of cue integration. In the next chapter we will examine a more general question: whether the human nervous system actually performs in accordance with Bayesian theory when the demand of perceptual fusion is loosened in a simple audio-visual perceptual task.

Chapter 3

Testing a Bayesian Model on an Auditory-Visual Illusion

Recently, it has been shown that visual perception can be radically altered by signals of other modalities. For example, when a single flash is accompanied by multiple auditory beeps, it is often perceived as multiple flashes. This effect is known as the sound-induced flash illusion.

In order to investigate the principles underlying this illusion, we developed an ideal observer (derived using Bayes rule), and compared human judgments with the ideal observer for this task. The human observers' performance was highly consistent with the ideal observer in all conditions ranging from no interaction, to partial integration, to complete integration, suggesting that the rule used by the nervous system for when and how to combine auditory and visual signals is statistically optimal. Our findings show that the sound-induced flash illusion is an epiphenomenon of this general, statistically optimal strategy.

3.1 Introduction

Situations in which an individual is exposed to sensory signals in only one modality are the exception rather than the rule. At any given instant, the brain is typically engaged in processing sensory stimuli from two or more modalities, and in order to achieve a coherent and ecologically valid perception of the physical world, it must determine which of these temporally coincident sensory signals are caused by the same physical source/event, and thus should be integrated into a single percept.

Spatial coincidence of the stimuli is not a very strong or exclusive determinant of cross-modal binding: information from two modalities may get seamlessly bound together despite large spatial inconsistencies (e.g., ventriloquism effect), while spatially concordant stimuli may be perceived as separate entities (e.g., someone speaking behind a screen does not lead to the binding of the voice with the screen). This is not surprising considering the relatively poor spatial resolution of auditory, olfactory, and somatosensory modalities. The degree of consistency between the information conveyed by two sensory signals, on the other hand, is clearly an important factor in determining whether the cross-modal signals are to be integrated or segregated.

Previous models of cue combination [Bulthoff and Mallot, 1988, Knill, 2003, Landy et al., 1995, Yuille and Bulthoff, 1996, Alais and Burr, 2004] have been successful in accounting for a variety of cross-modal tasks including visual-haptic [Masaro, 1998, Ernst and Banks, 2002], sensory-motor [van Beers et al., 1999], and visual-proprioceptive [Ghahramani et al., 1988] tasks. However, these models [Masaro, 1998, Ghahramani et al., 1988, Shams et al., 2000] have all focused exclusively on conditions in which the signals of the different modalities get completely integrated (or appear so due to the employed paradigms that force subjects to only report one percept, and thus not revealing any potential conflict in percepts). Therefore, the previous models are unable to account for the vast number of situations in which the signals do not get integrated or only partially integrate [Shams et al., 2002].

3.2 The sound induced flash illusion

The sound-induced flash illusion [Jordan, 2004, Falchier et al., 2002] is a psychophysical paradigm in which both integration and segregation of auditory-visual signals occur depending on the stimulus condition. When one flash is accompanied by one beep (i.e., when there is no discrepancy between the signals), the single flash and single beep appear to be originating from the same source, and are completely fused.

However, when one flash is accompanied with 4 beeps (i.e., when the discrepancy is large), most often they are perceived as emanating from two separate events, and the two signals are segregated, i.e., a single flash and four beeps are perceived. If the single flash is accompanied by two beeps (i.e., when the discrepancy is small), the single flash is often perceived as two flashes and on these illusion trials, the flashes and beeps are perceived as

having originated from the same source, i.e., integration occurs in a large fraction of trials. When a single flash is accompanied by three beeps, on a fraction of trials the single flash is perceived as two flashes while three beeps are perceived veridical. These trials would exemplify conditions of partial integration, in which the visual and/or auditory percepts are shifted towards each other but do not converge.

Therefore, the sound-induced flash illusion offers a paradigm encompassing the entire spectrum of bi-sensory situations. Because signals are not always completely integrated, previous models of cross-modal integration cannot account for these effects. Therefore, we developed a new model, which is a generalization of some previous models, in order to be able to account for the situations of segregation and partial integration, as well as complete integration. The model is an ideal observer and in contrast to previous models of cue combination, it does not assume one source for all the sensory signals (which would enforce integration), and instead it assumes one source for the signal in each modality.

However, the sources are not taken to be statistically independent, and thus, the model allows inference about both cases in which separate entities have caused the sensory signals, and cases in which sensory signals are caused by one source. The model uses Bayes rule to make inference about the causes of the various sensory signals.

We presented observers with varying combinations of beeps and flashes, and asked them to report the perceived number of flashes and beeps in each trial. We then compared the human judgments with those of the ideal observer.

3.3 Methods

3.3.1 Stimuli

The visual stimulus consisted of a uniform white disk extending 1.5 degrees of visual field at 12 degrees eccentricity below the fixation point (Figure 3.1a), flashed for 10 ms on a black computer screen one to four times. The auditory stimulus was a 10 ms long beep with 80 dB SPL and 3.5 kHz frequency, also presented one to four times. A factorial design was used in which all combinations of 0-4 flashes and 0-4 beeps (except for the no flash no beep combination) were presented, leading to a total of 24 conditions. The SOAs of flashes and beeps were 70 ms and 58 ms, respectively (Figure 3.1b). These specific SOAs were chosen because of certain constraints (e.g., frame rate, obtaining a strong illusion in the

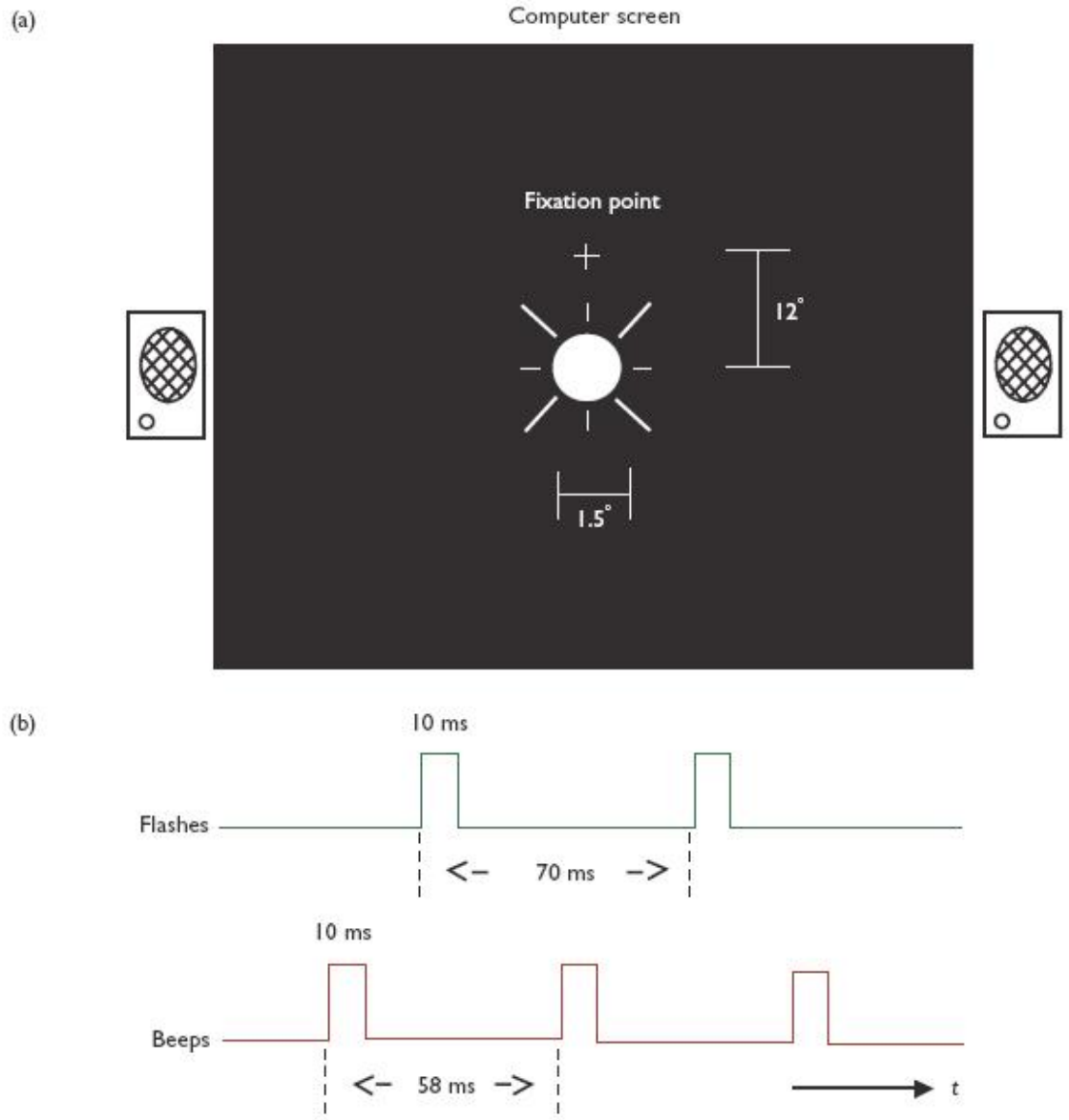


Figure 3.1: The spatio-temporal configuration of stimuli. (a) The spatial configuration of the stimuli. The visual stimulus was presented at 12 degrees eccentricity below the fixation point. The sounds were presented from the speakers adjacent to the monitor and at the same height as the center of the visual stimulus. (b) The temporal profile of the stimuli in one of the conditions (2 flashes+3 beeps) is shown.The centers of the visual and auditory sequences were aligned in all conditions.

illusion conditions, and the smallest sound SOA which consistently is above flutter fusion threshold). The behavioral data is fairly robust to the exact visual and auditory SOAs. The relative timing of the flashes and beeps was set such that the centers of the flash and beep sequences were synchronous in order to maximize the time overlap between the two stimuli. Sound was presented from two speakers placed adjacent to the two sides of the computer monitor, at the height where the visual stimulus was presented, thus, localizing at the same location as the visual stimulus.

3.3.2 Procedure

Ten naive observers participated in the experiment. Observers sat at a viewing distance of 57 cm from the computer screen and speakers. Throughout the trials there was a constant fixation point at the center of the screen. The observer's task was to judge both the number of flashes s/he saw and the number of beeps s/he heard after each trial (these reports provide $P(X_A, X_V|A, V)$ as described below). The experiment consisted of 20 trials of each condition, amounting to a total of 480 trials, ordered randomly. There was a brief rest interval after each third of the experiment.

3.3.3 The Ideal Observer Model

We assume that the auditory and visual signals are statistically independent given the auditory and visual causes (see Figure 3.2). This is a common assumption, motivated by the hypothesis that the noise processes that corrupt the auditory and visual signals are independent. This conditional independence means that if the causes are known then knowledge about V provides no information about A , and vice versa, as the noises corrupting the two signals are independent. In the meantime, if the causes are not known, knowledge of V provides information about A, and vice versa [Rockland and Ojima, 2003].

The information about the likelihood of sensory signal A occurring given an auditory cause X_A is captured by the probability distribution $P(A|X_A)$. Similarly, $P(V|X_V)$ represents the likelihood of sensory signal V given a source X_V in the physical world. The priors $P(X_A, X_V)$ denote the perceptual knowledge of the observer about the auditory-visual events in the environment. In addition to the observers experience, the priors may also reflect hard-wired biases imposed by the physiology and anatomy of the brain (e.g.,

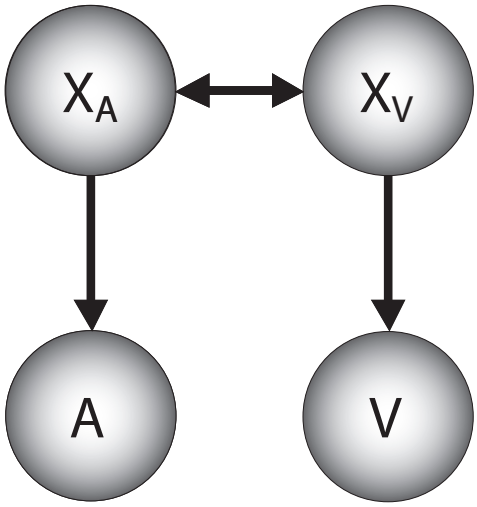


Figure 3.2: Graphical model describing the ideal observer. In a graphical model [Pearl, 1988], the graph nodes represent random variables, and arrows denote potential conditionality. The absence of an arrow represents direct statistical independence between the two variables. The bidirectional arrow between X_A and X_V does not imply a recurrent relationship; it implies that the two causes are not necessarily independent.

the pattern of interconnectivity between the sensory areas [Pearl, 1988, Clark and Yuille, 1990]), as well as biases imposed by the task or observers state, etc.

The graph in Figure 3.2 illustrates the two key features of the model. First, that there are two sources, X_A and X_V, for the two sensory signals A and V . This allows inference in both cases in which the signals A and V are caused by the same source and cases in which they were caused by two distinct sources; i.e., in contrast to the previous models, this model does not a priori assume that the signals have to be integrated. Second, in this model, X_V influences A only through its effect on X_A, and likewise for X_A and V . This corresponds to the assumption of independent likelihood functions, $P(A, V|X_A, X_V) = P(A|X_A)P(V|X_V)$. This is a plausible assumption motivated by the fact that either the two signals are caused by two different events in which case A would be independent of X_V (and likewise for V and X_A), or they are caused by one event in which case the dependence of A on X_V can be

captured by its dependence on X_A.

Given the visual and auditory signals A and V, an ideal observer would try to make the best possible estimate of the physical sources X_A and X_V, based on the knowledge $P(A|X_A)$, $P(V|X_V)$, $P(X_A, X_V)$. These estimates are based on the posterior probabilities, which can be calculated using Bayes rule, and simplified by the assumptions represented by the model structure (Figure 3.2), resulting in the following inference rule:

$$P(X_A, X_V|A, V) = \frac{P(A|X_A)P(V|X_V)P(X_V, X_A)}{P(A, V)} \tag{3.1}$$

This inference rule simply states that the posterior probability of events X_A and X_V is the normalized product of the single-modality likelihoods and joint priors. To relate this to the subjects' responses we assume that they were performing probability matching [Vulkan, 2000], i.e., that they directly reported their subjective posterior probabilities. In order to simplify calculations, we also assume that $P(A, V)$ has a uniform distribution. This, in turn, implies that $P(A)$ and $P(V)$ also have uniform distributions. Given a uniform $P(A)$, the auditory likelihood term is computed as follows

$$P(A|X_A) = \frac{P(X_A|A)P(A)}{\sum P(X_A|A)P(A)} = \frac{P(X_A|A)}{\sum P(X_A|A)} \tag{3.2}$$

(and likewise for $P(V|X_V)$). While the likelihood functions $P(A|X_A)$ and $P(V|X_V)$ are nicely approximated from the unisensory (visual-alone and auditory-alone) conditions, the prior probabilities $P(X_A, X_V)$ involve both sensory modalities and cannot be obtained from unisensory conditions alone.

3.3.4 Estimation of the joint priors

In most models, the priors are not directly computable. Hence, the prior distribution is parameterized and the parameters are tuned to fit the observed data (i.e., data to be predicted). Our experimental paradigm makes it possible for the joint priors to be approximated directly from the observed data, alleviating the need for any parameter tuning. The joint priors can be approximated by marginalizing the joint probabilities across all conditions, i.e., all combinations of A and V:

$$P(X_A, X_V) = \sum_{A,V} P(X_A, X_V|A, V)P(A, V). \tag{3.3}$$

Given a uniform $P(A, V)$, this leads to a normalized marginalization of the posteriors. Because this estimate requires marginalizing over all conditions, including auditory-visual conditions, we used the data from a different set of observers (the first half of participants) for estimating the joint priors using the above formula, and excluded those data from the testing process (the second half of participants). In other words, this data was only used for calculating the priors and discarded afterwards. Thus, the model remained predictive, not using any auditory-visual data for making predictions about the performance in the auditory-visual conditions. Although it may appear that the joint prior matrix introduces 24 free parameters in our model, it should be emphasized that this is not the case, as these parameters are not free. The parameters of the joint prior matrix are set using the observed data, however, they were not tuned to minimize the error between the model predictions and the data. Therefore, the model has no free parameters. To calculate the goodness of fit we use amount of explainable variance, $r^2 = 1 - var(data - model)/var(data)$, and the sum of the residuals, $sse = \sum(model - data)^2$.

3.4 Results

The data for visual-alone and auditory-alone conditions are shown in the first column and first row of Figure 3.3, respectively. Each panel represents one condition, i.e., a particular combination of the auditory and visual stimuli. To facilitate interpretation of the data, instead of presenting a 5x5 matrix of joint posterior probabilities for each condition only the two one-dimensional projections of the joint posteriors are displayed; i.e., in each condition, instead of showing 25 auditory-visual posteriors, 10 marginalized posteriors are shown (5 auditory, and 5 visual). As can be seen, the observers perform better in the auditory-alone conditions. The plots in rows and columns 2-5 of Figure 3.3 display the data for the auditory-visual conditions. As can be seen, the human observers performance is remarkably consistent with the ideal observer in all of the conditions ($r^2 = 0.93$, $sse = 0.900$), accounting for 160 data points (5 X_A and 5 X_V combinations at 16 conditions) with no free parameters.

Examining the various plots in Figure 3.3 reveals that only in conditions where the

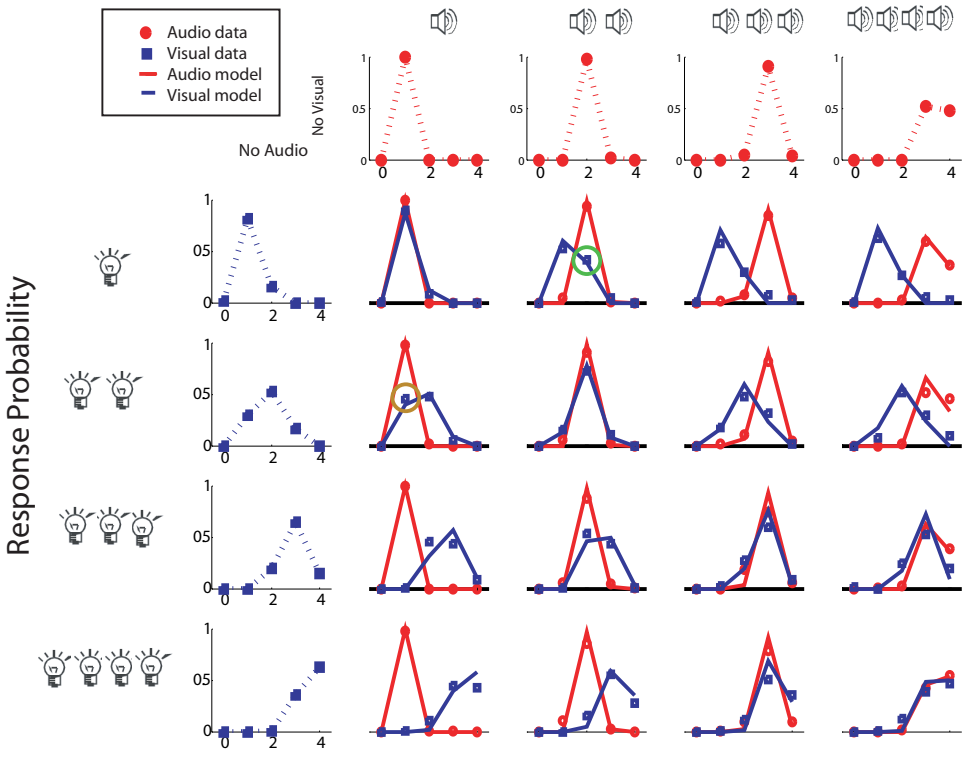


Figure 3.3: Comparison of the performance of human observers with the ideal observer. To facilitate interpretation of the data, instead of presenting joint posterior probabilities for each condition, only the marginalized posteriors are shown. The auditory and visual judgments of human observers are plotted in red circles and blue squares, respectively. Each panel represents one of the conditions. The first row and first columns represent the auditory-alone and visual-alone conditions, respectively. The remaining panels correspond to conditions in which auditory and visual stimuli were presented simultaneously. The horizontal axes represent the response category (with zeros denoting absence of a stimulus and 1-4 representing number of flashes or beeps). The vertical axes represent the probability of a perceived number of flashes or beeps. The data point, which is enclosed by a green circle, is an example of the sound-induced flash illusion, showing that in a large fraction of trials, observers perceived two flashes when one flash was paired with two beeps. The data point enclosed by a brown circle reveals an opposite illusion in which two flashes are perceived as one flash in a large fraction of trials in the 2 flashes+1beep condition.

visual and auditory stimuli are identical (i.e., the conditions displayed along the diagonal) do observers consistently indicate perceiving the same number of events in both modalities. In conditions in which the inconsistency between the auditory and visual stimuli is not too large, for instance, in 1flash+2beep condition or 2flash+1beep condition, there is a strong tendency to combine the two modalities, as indicated by highly overlapping auditory and visual reports. The high values along the diagonal in joint posterior matrices of these conditions (not shown here) confirm that indeed the same number of events were experienced jointly in both modalities in these conditions.

The integration of the auditory-visual percepts is achieved in these cases by a shift of the visual percept in the direction of the auditory percept. This occurs because the variance in the auditory-alone conditions is lower than those of the visual-alone conditions. In other words, because the auditory modality is more reliable it dominates the overall percept in these auditory-visual conditions. This finding is consistent with previous studies of cue-combination within [Landy et al., 1995, Yuille and Bulthoff, 1996, Jacobs, 1999, Battaglia et al., 2003] or across modalities [Masaro, 1998, van Beers et al., 1999, Ghahramani et al., 1988] in all of which the discrepancy between the two cues is small, and the percept is dominated by the cue with lower variance (or higher reliability). The large fraction of trials in which the observers report seeing two flashes in the 1flash+2beeps condition corresponds to the sound-induced flash illusion [Shams et al., 2002].

In conditions in which the discrepancy between the number of flashes and beeps is large (e.g., 1flash+4beeps or 4flashes+1beep), the overlap between the auditory and visual percepts is significantly smaller, indicating a considerably smaller degree of integration and larger degree of segregation. These conditions of large discrepancy have not been investigated by the previous studies of within-modality or multi-sensory cue combination, and no general model has been previously offered that would account, in a coherent manner, for both the situations of small and large discrepancy.

Next, we investigated the possibility that the Bayesian model of equation (3.1) is overly powerful and capable of predicting any dataset. We shuffled the obtained human observer posterior probabilities $P(X_A, X_V|A, V)$ in each auditory-visual condition leading to a new dataset that was identical to the human data in its overall content, though randomized in order. We applied our model to this data set. The model predictions did not match the shuffled data, even when we did not divide the dataset into two halves and instead computed

the priors from the same set that we generated the predictions for (r^2 =-1.51, $sse = 22.13$). We obtain qualitatively similar results regardless of the made-up distribution used. This finding strongly suggests that the predictions of the proposed ideal observer are distinctly consistent with the human observers data and not with any arbitrary data set.

3.5 Discussion

Altogether, these results suggest that humans combine auditory and visual information in an optimal fashion. These findings are consistent with the report of optimality of human visual-haptic integration [Masaro, 1998], as well as a recent study showing that a non-uniform prior [Knill and Pouget, 2004] is needed to account for the ventriloquism effect. Our results extend these earlier findings by showing that the optimality of the human performance is not restricted to situations in which the discrepancy between the two modalities is minute and the two modalities are completely integrated. Indeed, it can be shown that many earlier models of cue combination are special cases of the model described here

The ideal observer model presented here differs from the previous models of cue combination, which have employed maximum likelihood estimation in two important ways. First, as opposed to previous models (which assume one cause for all signals) our model allows a distinct cause for each signal. This is a structural difference between the present model and all the previous models, and is the reason why the multi-sensory paradigm in the present study is beyond the scope of previous models. The assumption of one cause makes previous models unable to account for a vast portion of the present data, where the visual and auditory information are not integrated, i.e., all the trials in which subjects report different visual and auditory percepts.

Second, the previous models did not include any prior probability of events, which is equivalent to assuming a uniform prior distribution. In the present model prior probabilities are not assumed to be uniform. In order to examine the importance of the priors in accounting for the data, we tested the model using a uniform prior. The goodness of fit was considerably reduced ($r^2 = 0.79$, $sse = 2.60$) indicating that in this task the priors depart significantly from uniform distribution, and are therefore necessary for accounting for the data.

The findings of this study suggest that the brain uses a mechanism similar to Bayesian

inference [Knill and Pouget, 2004] for deciding whether or not, to what degree, and how (in which direction) to integrate the signals from auditory and visual modalities, and that the sound-induced flash illusion can be viewed as an epiphenomenon of a statistically optimal computational strategy.

We made a few assumptions here that deviate from a perfectly Bayesian model. For example we assumed that subjects perform their decision by sampling their posterior, thereby directly reporting their posterior. This also allowed us to explain the variability in the data, which should not have been there if subjects simply chose the maximum or mean of the posterior. However this is not an optimal strategy given a typical utility function.

However, it is known that humans do perform in accordance with a 'probability matching ' approach under certain conditions [Vulkan, 2000], and this decision rule has been used in other Bayesian psychophysics studies [Mamassian and Landy, 2001]. One argument for why subjects would use this, is that there may be long-term advantages to sampling a posterior, such as learning [Sabes and Jordan, 1996], if the gain/loss of the current task is not too high. In a later chapter we will see though that there are better approaches to explaining the variance in subjects responses that still allow us to estimate the parameters in the model.

For simplicity, we don't take any account of the boundary effects, although they can clearly have an effect. If a subject perceives 5 flashes, but is forced to respond 4, this can certainly create asymmetry in the distributions.

That we assume that subjects 'report' their posteriors, also allows us to estimate the prior from the marginalized posterior. If we were to abandon this decision rule, clearly this relationship would no longer hold. Furthermore we assume that $P(A, V)$ is uniform in order for us to to be able to calculate the prior. We may present the stimuli uniformly, but there is no way for us to test whether it actually is perceived as flat.

Further chapters will discuss some of these assumptions.

3.6 Summary

A full-prior Bayesian model is able to explain a large amount of the variance for an auditory-visual illusion. However in the process we had to make some assumptions that are clearly not in accordance with a fully optimal model, especially with regard to the utility function and decision rule.

Another problem with the full model we have here presented is the unspecified prior $P(X_A, X_V)$. We have in the preceding assumed that the prior could take any form and allowed the marginalized posterior to specify it. However we would assume that there is an underlying specific shape for this distribution based on regularized principles. Is there a structured way to constrain this prior? We will see that using the concept of causality and assuming that the brain uses this, allows us to nicely perform such constraints.

Chapter 4

Causal Inference

The problem of causality has long been debated, but has only during the 20th century been analyzed from a statistical viewpoint. We present the theory behind and show how it can be applied to subject's judgments.

4.1 Causality

Causality in its modern description is most well known from the writings of David Hume [Hume, 1777]. Given the collision between a stationary and a moving ball, how can we claim that the moving ball caused the previously stationary ball to start moving? Does it follow logically? Obviously not as we can easily imagine counter examples that do not involve the ball to start moving. Can we then make the inference from just observing a single occurrence? The concept of cause and effect implies a necessity of the effect, that one could not deduce from a single observation (according to Hume).

We have to rely on induction based on multiple observations of a cause and an effect. But induction has problems too: having only observed white swans does not preclude us from ever observing black swans. The past does not guarantee the future.

Logically (according to Hume, again) we can therefore never make statements about causal relationships.

More stringently using a more modern formulation (here following [Holland, 1986]), in order for us to claim that A causes B_1 to change state to B_2 we need to observe the effect

$$Effect = B_1 - B_2,$$

that is, the difference between whether we had performed action A or not. However, as we either perform A or we do not perform A, we can never observe both B_1 and B_2 and can therefore, in theory, never determine if A did indeed cause the shift in B. As we shall see, unsurprisingly, there are ways to avoid this problem in certain situations.

4.2 Solutions to the problem of causality

One solution to the problem above comes from what is referred to as the scientific solution. If we have observed the motionless ball for an extended time, before the contact with the moving ball, we feel justified in knowing the state B_1, what were to happen if the moving ball had not touched it. The effect therefore simply becomes the difference in time: [after the collision] minus [before the collision].

This example works for such simple setups that are (nearly) deterministic and which have no natural temporal development, but how do we deal with more complex situations, say the application of some new drug towards treatment of a certain disease? Observing the subject before and after treatment of the drug will not inform us as to whether the drug causes the subject's condition to improve. We are dealing with a time varying non-linear system; for all we know the subject could be getting better/worse whether or not we had applied the drug. The statistical solution is to have two subject groups; one that receives the drug and one that only receives a placebo. Assuming we have well-matched and large enough groups, statistics allows us to state with a given certainty (p-value) whether the drugs are efficient. In order for there to be an effect we at least require that $P(E|drug) > P(E|\neg drug)$, i.e., that the probability of an effect is larger for the group that received the drug than the group that received a placebo.

This solution is similar to the scientific solution in that it utilizes an assumption of transferability from the average case to the general case, in the scientific solution we average over time before and after, in the statistical solution we average over two subject groups.

Of course in order for us to be able to compare two population groups, they need to be well matched in 'all relevant' aspects. But what are relevant aspects? Age, ethnicity, hair color? Often this is not known and we hope that by having a large enough a sample we can average over any other effects.

To believe that causality is a trivial problem would therefore surely be a mistake.

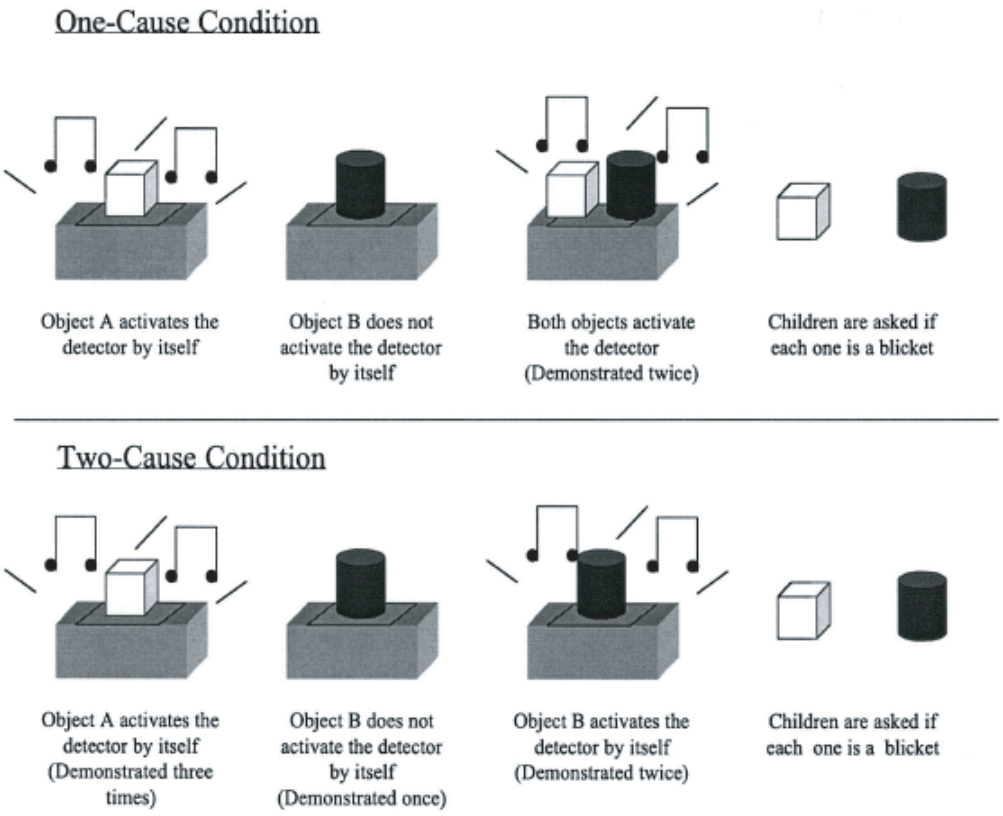


Figure 4.1: Procedure used by Gopnik et al. Several different conditions were used including the one-cause and two-cause condition depicted here. Notice that the setup was not entirely deterministic, as witnessed in the two-cause condition. (From [Gopnik et al., 2004].)

However, there is no doubt that we employ this principle on a daily basis (in simpler situations), and quite successfully. We are able to navigate the world and perform complex tasks in part due to the fact that we can rely on certain effects to follow certain causes. If I were to walk out in front of a car I would certainly feel the detrimental effects of the collision.

4.3 Cognitive examples of causal inference

A series of experiments have explicitly examined how humans are able to perform such inference. Gopnik et al. [Gopnik et al., 2001] presented children age 2-4 with a 'blicket detector'. Wooden blocks of different colours would be placed in different combinations on top of a machine that would light up whenever one of the boxes was a 'blicket'. The

subjects were not able to influence which blocks would be placed on the detector, but could only observe. Nevertheless the children were quite adept at identifying the 'blicket', i.e., the block that would be the cause of the activation of the machine, even through watching only a few trials. Traditional theories have focused on bottom-up approaches; subjects would gather enough data to be able to ascertain statistical correlations and based on that build up a model for the causal structure of the task. However with just a few trials to base a decision upon, there is clearly not enough data available to do this. Alternative theories assume [Griffiths and Tenenbaum, 2005] that subjects already possess a certain expectation of the structure of the causes and effects in the world and use Bayesian inference to weed out the different conflicting hypothesis. In this viewpoint subjects are performing hypothesis selection.

Given the data pairs (e.g., block A and B activate detector E) the model tries to infer which of the blocks is the actual cause. As an example, imagine the subject is being shown the setup in Figure 4.1a. Before the experiment several possible hypothesis were available, A could be a blicket, B could be a blicket or both could be blickets. Given the data, D, Bayes' rule gives a simple way of estimating the probability of a hypothesis h to be correct:

$$P(h|D) = P(D|h)P(h)/P(D). \tag{4.1}$$

If we wish to compare hypotheses, say h_1 and h_2, we can take their ratio

$$P(h_1|D)/P(h_2|D) = P(D|h_1)P(h_1)/P(D|h_2)P(h_2) \tag{4.2}$$

which is just the ratio of their likelihoods and posteriors. A ratio larger than 1 favors hypothesis 1, otherwise hypothesis 2 is preferable. If their prior probabilities are equal, the ratio is just equal to the ratio of their likelihoods. In the current example the likelihood clearly favors the hypothesis that block A is the blicket.

Such a Bayesian approach is much better able to explain the behavior of subjects than any correlation-based model [Griffiths and Tenenbaum, 2005].

In another study, subjects were told about a new fictional explosive material 'Nitro X' [Griffiths et al., 2004]. 'Nitro X' is highly unstable to the point where gently tapping a can filled with it will cause it to explode. If it is placed adjacent to any other cans of

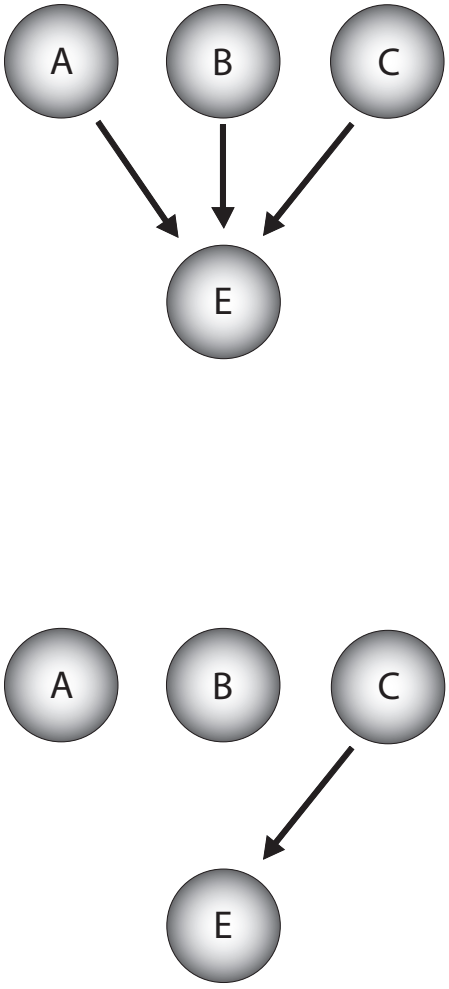


Figure 4.2: Bayesian model to explain causal inference. a) Subjects are presented with several examples of blocks (A, B, C) and shown whether the combination activates the 'giblet detector.' b) Given this information (effect E) they have to infer which of the blocks is the cause.

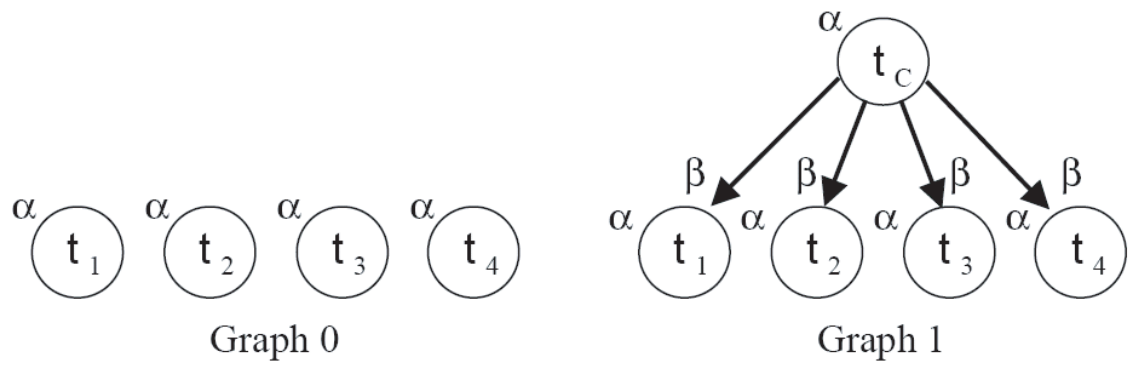


Figure 4.3: Two causal models used to describe the 'Nitro X' experiment in Griffiths et al. [Griffiths et al., 2004]

Nitro, one can exploding will subsequently cause neighboring cans to explode too. Further, subjects were informed that cans were capable of exploding spontaneously without any outside interference. After thus being familiarized with 'Nitro X', subjects were shown animated movies of a row of cans exploding, in some cases starting with one can generating a chain reaction of exploding cans, but in other movies the cans would explode simultaneously, inconsistent with a chain reaction. Subjects would then be asked to indicate whether one can's explosion was the cause of the others explosion or whether they all exploded due to some hidden cause. As the number of canisters exploding simultaneously increased, the subjects were more likely to report that a hidden cause must be affecting them.

Again, a model that assumes that subjects learn the structure by looking for statistical correlations would be very hard pressed to explain the data, as subjects perform this after only observing single occurrences. However a Bayesian model that assumes that subjects already have some prior structure imposed on the data has a much better performance. A graphical representation of two conflicting causal models is shown in Figure 4.3. One hypothesis is that each can has a small chance at each time step to explode, taking the other cans with it in the process. The other hypothesis assumes some hidden cause, that directly causes the all canisters to explode simultaneously. As in the previous example, a ratio of the posterior probability of the two hypothesis can be computed, to indicate which one is the most likely.

4.4 Summary

Given very limited amounts of information, humans even at a very young age are capable of making inferences about data sets for which a full statistical analysis is simply not possible. This implies some type of prior structure or knowledge is utilized, and an ideal way to think of this is therefore to use a Bayesian framework for modeling the data.

The experiments described above have all involved cognitive decisions; in the next chapter we will look at how causal modeling can be applied to perceptual problems and be used to restrain our prior from Chapter 3.

Chapter 5

Testing Causal Inference

As discussed, to accurately perceive the world we need to integrate signals from multiple senses. A major limitation of the current understanding of this integration is that it does not incorporate the realization that only signals that have the same cause should be integrated. Here we used causal inference to calculate the likelihood of a common cause to obtain optimal estimates of the position of the source(s). We accurately predict the nonlinear integration of cues by human observers in psychophysical experiments. The capacity to infer causal structure is thus not limited to high-level cognition; it is performed continually and effortlessly in perception.

5.1 Introduction

When we see a friend talk we perceive the voice as coming from the mouth, but this percept can be fooled: a good puppeteer makes us perceive his voice as coming out of a puppet's mouth. In both cases, we infer a common cause to explain the correlation between the mouth's movement and voice's sound. Numerous psychophysical studies have investigated how two cues are integrated into a common estimate of the source's position [Pick et al., 1969, Wallace et al., 2004, Hairston et al., 2003, Thomas, 1941, Thurlow and Jack, 1973]. These position estimates are often consistent with an optimal Bayesian strategy [Jacobs, 1999, van Beers et al., 1999, Ernst and Banks, 2002, Alais and Burr, 2004].

However, the brain must solve a harder problem than these models do: we are often surrounded by several sources of sensory stimulation. For example a car alarm may go off while someone is speaking. Given any two signals from different modalities, such as vision and audition, how does the brain decide whether they have a common cause or two indepen-

dent causes – and hence whether they should be integrated or processed separately? Here we formalize these problems and show that we can accurately predict human performance over a range of cue combination tasks.

5.2 Causal inference

We model situations in which subjects are presented with simultaneous auditory and visual stimuli, and are asked to report their location(s). If the visual and the auditory stimuli have a common cause (Figure 5.1a), subjects could use visual cues to improve the auditory estimate and vice versa. Traditional models of cue combination assume that this is always the case [Jacobs, 1999, van Beers et al., 1999, Ernst and Banks, 2002, Alais and Burr, 2004].

However, in the real world we are usually surrounded by many sights and sounds. Therefore we cannot simply combine all signals into a joint estimate; we must infer which signals have a common cause and only integrate those. Specifically, for any pair of visual and auditory stimuli, we should consider not only the hypothesis that they have a common cause (Figure 5.1a) but also the alternative possibility that they are unrelated (Figure 5.1b). An ideal observer model based on causal inference needs to estimate if there is a common cause.

Two factors need to be considered: the prior belief in a common cause and the positions of the perceived auditory and visual stimuli (see Methods below for details). As the causal inference model can never be certain that there is one versus two causes, it will consider both possibilities weighted by their relative probabilities. This model depends on only 4 parameters characterizing knowledge about the situation and the observer's sensory systems: on the precision of vision (σ_{vis}) and audition (σ_{aud}); on knowledge the observer has about the world, in particular how much the observer expects that stimuli are more likely to be straight ahead (σ_{pos}); and, how likely a *priori* the observer assumes that there is a single cause versus two causes (p_{common}). These 4 parameters are obtained by fitting human behavior. We were thus able to make predictions of human behavior in psychophysical experiments.

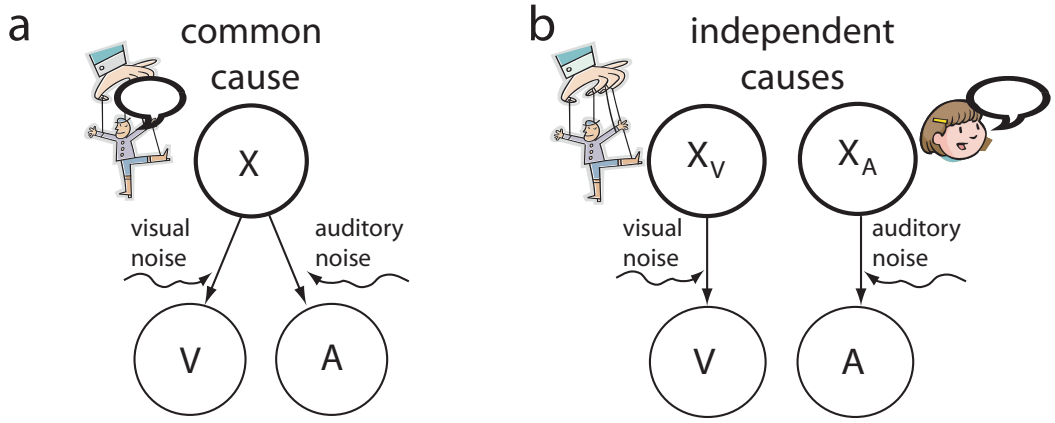


Figure 5.1: The causal inference model a) One cause can be responsible for both cues. In this case the visually perceived position V will be the common position x perturbed by visual noise with width σ_{vis} and the auditory perceived position will be the common position perturbed by auditory noise with width σ_{aud}. b) Alternatively, two distinct causes may be relevant, decoupling the problem into two independent estimation problems. The causal inference model infers the probability of causal structure a) versus the causal structure b) and then derives optimal predictions from this.

5.3 Ventriloquist Illusion

To test the causal inference model, 20 subjects made a series of perceptual decisions. On each trial, observers were presented with either a brief visual stimulus in one of five locations along the horizontal direction in the frontoparallel plane, or a brief sound at one of the same five locations, or both simultaneously These trials were presented in a pseudo-random order. The task of the subjects was to report the location of the visual stimulus as well as the location of the sound in each trial using two button presses during each trial (Figure 5.2a).

Methods

5.3.1 Experimental paradigm.

Twenty naive observers (undergraduate students at Caltech, ten male) participated in the experiment. Observers were seated at a viewing distance of 54 cm from a 21-inch monitor. In each trial, observers were asked to report both the perceived visual and auditory positions using keyboard keys 1 through 5, with 1 being the leftmost location and 5 the rightmost.

Visual and auditory stimuli were presented independently at one of five positions. The five locations extended from 10° to the left of the fixation point to 10° to the right of the fixation point along a horizontal line 5° below the fixation point, at 5° intervals.

Visual stimuli were 35 ms presentations of Gabor wavelets of high-contrast extending 2° on a background of visual noise. Auditory stimuli, synchronized with the visual stimuli in the auditory-visual conditions, were presented through a pair of headphones and consisted of 35 ms white noise. The sound stimuli were filtered through a Head Related Transfer Function (HRTF), measured individually for each subject, using methods similar to those described by http://sound.media.mit.edu/KEMAR.html (see Section C.4 for details). The HRTFs were created to simulate sounds originating from the five spatial locations in the frontoparallel plane where the visual stimuli were presented.

5.3.2 Generative model

We assume that for every near coincidence of cues there is a probability p_{common} that two cues have a common cause. Outside of the experiment this will not be constant but depend on temporal delays, visual experience, context, and many other factors. In the experiments we consider, all these factors are held constant so we can use a fixed p_{common}. We assume that the visual and the auditory signal are corrupted by unbiased Gaussian noise of standard deviations σ_{vis} and σ_{aud}, respectively. We assume that positions are drawn from a Gaussian distribution with a width σ_{pos}. We assume that people try to minimize the expected mean squared error of their pointing response, in contrast to Chapter 3 where we assumed that subjects used probability matching. We thus obtain a mixture model.

5.3.3 Inference

Inferring if there is a single or two causes is a Bayesian model decision problem. The probability that the two cues have a common cause depends on the visually and auditory perceived positions:

$$P(common|V,A) = \frac{p_{common}P(V,A|common)}{p_{common}P(V,A|common) + (1-p_{common})P(V,A|\neg common)} \quad (5.1)$$

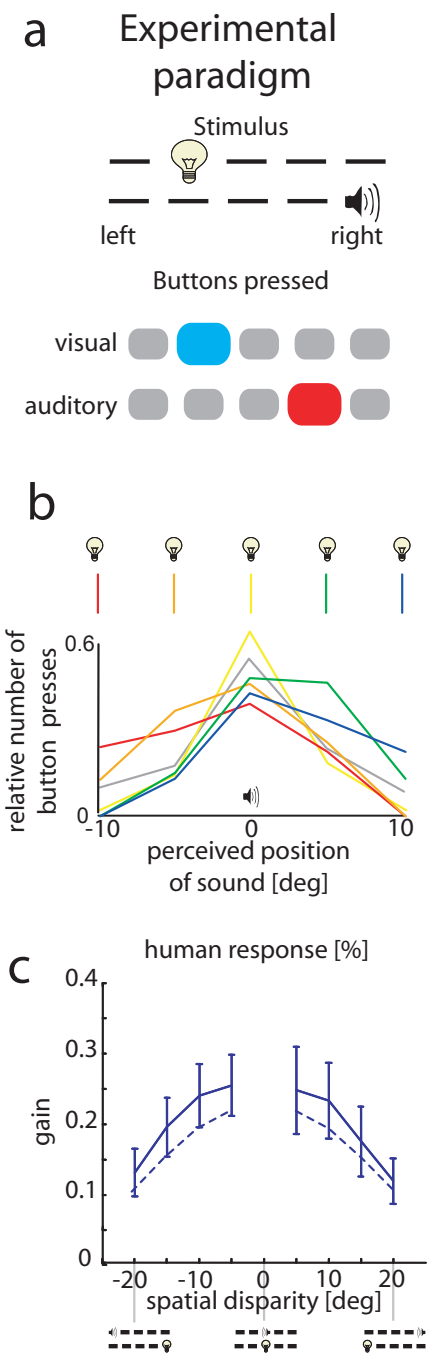


Figure 5.2: Combination of visual and auditory cues. a) The experimental paradigm is shown. In each trial a visual and an auditory stimulus is presented simultaneously and observers report both the position of the perceived visual and the position of the perceived auditory stimuli by button presses. b) The influence of vision on the perceived position of an auditory stimulus in the center is shown. Different colors correspond to the visual stimulus at different locations (sketched in warm to cold colors from the left to the right). The unimodal auditory case is shown in gray. c) The average auditory gain $((x_{est} - A)/(V - A)$, i.e., the influence of deviations of the visual position on the perceived auditory position is shown as a function of the spatial disparity (solid lines) along with the model prediction (dashed lines).

The optimal solution is:

$$\hat{x} = P(common|V,A)\hat{x}_{common} + (1 - P(common|V,A))\hat{x}_{\neg common} \tag{5.2}$$

From the Bayesian literature [Gharahmani, 1995] we know the optimal solutions $\hat{x}_{common}$ and $\hat{x}_{\neg common}$, as given by equations (2.6) and (2.12) and $P(V,A|common)$ is found by integration (see Appendix C.1.2). We are thus able to calculate the optimal estimate and the expected uncertainty given our assumptions. This model can be shown to be a restriction of the Bayesian model presented in Chapter 3, with a specific shape of the prior (see appendix C.3).

We choose p_{common}, σ_{pos}, σ_{vis}, and σ_{aud} to maximize the probability of the experimental data given the model. For the group analysis we obtain these 4 parameters from a dataset obtained by pooling together the data from all subjects. To obtain the histograms plotted in Figure 5.3 we simulate the system for each combination of cues for 10,000 times. To calculate the goodness of fit we use $r^2 = 1 - var(data - model)/var(data)$ and $sse = \sum(model - data)^2$. The optimization was done by minimizing the residuals, sse.

5.3.4 Results

We found that the visual stimulus influences the estimation of the auditory stimulus when the auditory stimulus is held at a constant location (straight ahead, Figure 5.2b). In this case, subjects clearly base their estimate of the auditory position on both visual and auditory cues. This finding is predicted by any cue combination strategy [Landy et al., 1995, Deneve et al., 2001], as the potential of a common cause generally implies an interaction between the cues.

The auditory position, however, clearly differs from the visually perceived position. Therefore, the combination of cues is not full, in contrast to the predictions of previous models [Jacobs, 1999, van Beers et al., 1999, Ernst and Banks, 2002, Alais and Burr, 2004]. Both the traditional model and the causal inference model correctly predict that the bimodal estimate will be more precise than the unimodal estimate (Figure 5.2b, gray). Human performance reveals a complicated pattern of partial combinations which is well predicted by the causal inference model (Figure 5.3). Moreover, the influence of vision on audition is much larger than the influence of audition on vision because visual perception in this

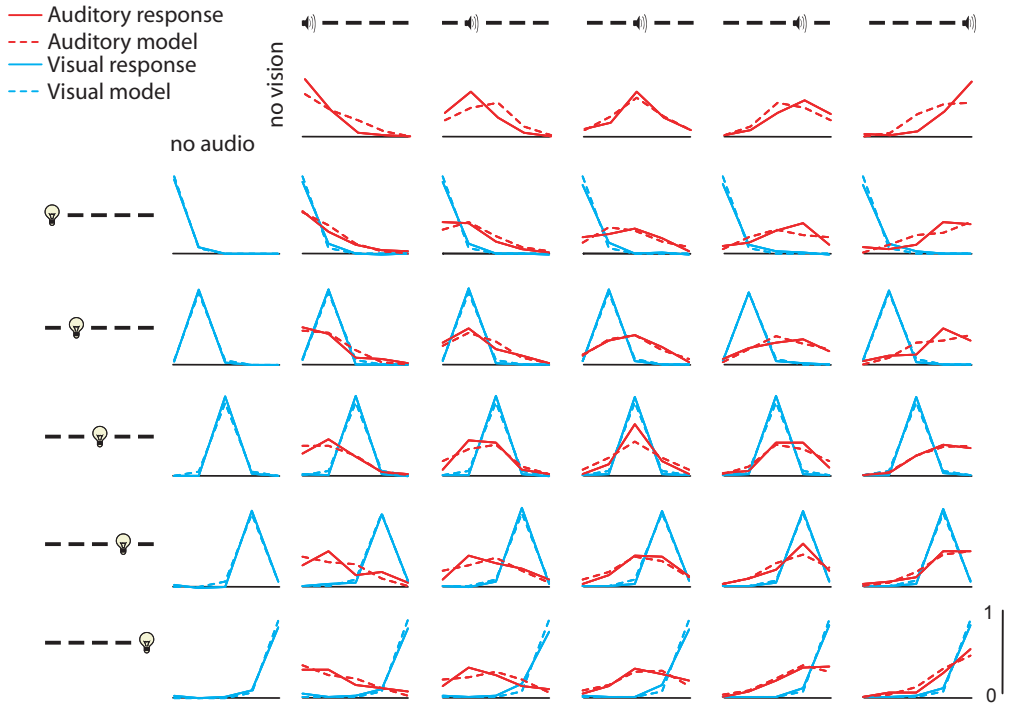


Figure 5.3: The average human observer responses (solid lines) are shown along with the predictions of the ideal observer (broken lines) for each of the 35 stimulus conditions. These plots show how often, on average, each button was pressed in each of the conditions.

experiment is much more precise than auditory perception.

The model explains 97% of variance ($sse = 0.54$) of the complicated pattern of human response probabilities as a function of visual and auditory position. To determine if the model's performance was due to overfitting the data, we used 10% of randomly chosen trials to predict 10% of the other trials and still found that the model accounts for 95±1% of the variance. Thus the model does not overfit the data. Importantly, when we take data from a recent experiment [Hairston et al., 2003] to estimate visual and auditory precision, and assume that people use the prior that is dictated by the distribution during the experiment, and that a common cause is a *priori* as likely as two independent causes, we still predict 92% of the variance ($sse = 1.41$) with a model without any free parameters. The model is thus highly robust in explaining the human data.

We also found that the auditory gain, the influence of vision on audition, decreases with

increasing spatial disparity as predicted by the model (Figure 5.2c). The further apart the two stimuli are, the more likely it is that the model infers that there are two causes. In the two-cause case, there is no combination of the cues, and for that reason the influence of the visual stimulus on the auditory perception decreases. A model in which no combination happens at all ($p_{common} = 0$) still explains 90% of the variance because the gain is relatively low ($sse = 1.25$). However, this model cannot explain the observed gains (Figure 5.2c) as it predicts a gain of zero ($p < 0.0001$ t-test).

A traditional full combination model [Jacobs, 1999, van Beers et al., 1999, Ernst and Banks, 2002, Alais and Burr, 2004] can only explain 56% of the variance, but predicts a gain of one, which is ruled out by the gain data (Figure 5.2c) ($p < 0.0001$ t-test). Neither traditional nor trivial no-combination models can explain the data, only causal inference can explain the observed patterns of partial combination.

To analyze how subjects vary with respect to their perceptual strategies we next fitted the parameters individually to each subject and reported means and standard errors over all subjects. We found the visual system to be relatively precise ($\sigma_{vis} = 1.35 \pm 0.2^\circ$) and the auditory system to be much less precise ($\sigma_{aud} = 8.85 \pm 1.0^\circ$). We found that people have a modest prior estimating stimuli to be more likely to be straight ahead ($\sigma_{pos} = 10.5 \pm 2.3^\circ$). The prior probability of perceiving a common cause is relatively low ($p_{common} = 0.26 \pm 0.05$) explaining the small gains observed (Figure 5.2c). We can thus rule out both full combination and complete separation processing strategies. In summary, the causal inference model provides precise predictions of the way people combine cues.

5.4 Causal percept

While the causal inference model provides an impressive account of the cue combination experiments described above it makes a prediction that goes well beyond just the estimates of positions and any previous models. If people infer the probability of common cause then it should be possible to ask them if they perceive a common versus two causes while also asking them for the localization of a target. Indeed in published experiments this was already done [Hairston et al., 2003, Wallace et al., 2004].

We thus compared the model's performance to data from published cue combination experiments. These experiments differed in a number of important respects from our ex-

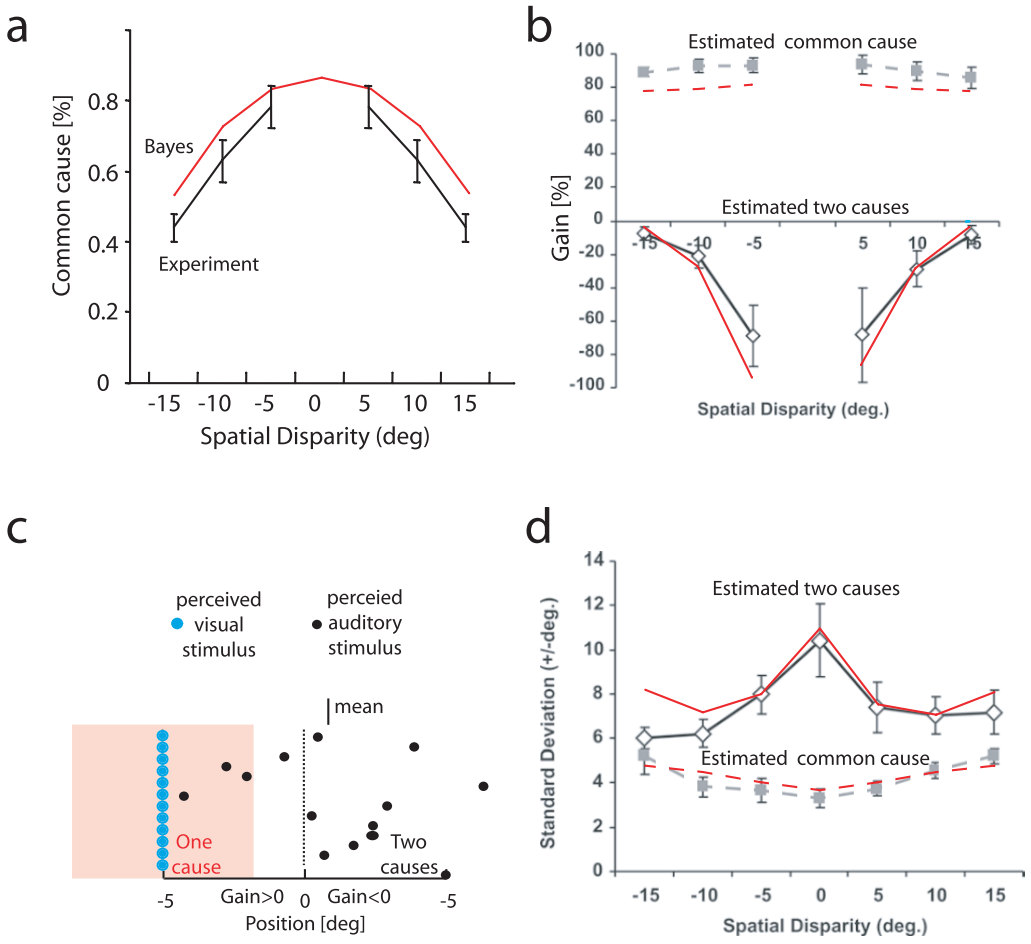


Figure 5.4: Reports of causal inference. a) The relative frequency of subjects reporting one cause (black) is shown (reprinted from [Wallace et al., 2004]) with the prediction of the causal inference model (red). b) The gain, i.e., the influence of vision on the perceived auditory position, is shown (gray and black). The predictions of the model are shown in red. c) A schematic illustration explaining the finding of negative gains. Blue and black dots represent the perceived visual and auditory stimuli, respectively. In the pink area people perceive a common cause. d) The standard deviation of the pointing direction of humans (black) and the model (red) is shown. The solid and broken lines correspond to trials in which a common cause or distinct causes were reported, respectively.

periments. Observers were asked to report their perception of unity (i.e., whether the two stimuli have a common cause or two independent causes) in each trial. Only the location of the auditory stimulus was probed. Subjects pointed towards the location of the auditory stimulus instead of choosing a pre-specified button to indicate the position. This procedure may lead to higher precision of the responses.

5.4.1 Data analysis

In these experiments, pointing is used to report the perceived position. In such cases there may be additional noise due to a misalignment of the cursor relative to the intended position of the cursor. To model this we introduce one additional variable, motor-noise. We assume that motor-noise corrupts all reports, is additive, and is drawn from a Gaussian with width σ_{motor}. We estimate the relevant uncertainties as follows: In both auditory and visual trials the noise will have two sources, motor noise and sensory noise. We assume that visual only trials are dominated by motor noise, stemming from motor errors and memory errors and that the noise in the visual trials is essentially exclusively motor noise. From other experiments ([Hairston et al., 2003] Figure 2) where movements are made towards unimodally presented cues, we obtain $\sigma_{motor} = 2.5^\circ$ and from the same graph, because variances are added linearly, we obtain $\sigma_{aud} = \sqrt{8^2 - 2.5^2}^\circ = 7.6^\circ$. These parameters were not tuned. The other two parameters σ_{pos} and p_{common} were obtained as the ML estimate. The same parameter values are used for all graphs in Figure 5.4.

5.4.2 Results

The outcome of these experiments [Hairston et al., 2003, Wallace et al., 2004] indicate that the closer the visual stimulus is to the auditory stimulus, the more often do people perceive them as having a common cause (Figure 5.4a). The causal inference model predicts the same pattern for the probability of a common cause. Even if the two stimuli are close to one another, on some trials the noise in the perception of the auditory stimuli will sometimes lead to the perception of there being distinct causes. The model explains 98% of the variance in human performance (Figure 5.4a) and thus accurately models the human perception of causality.

We next examined how the perception of common versus distinct causes affects the strategy used to estimate the position of auditory stimuli. When people perceive a common

cause they point to a position that is on average very close to the position of the visual stimulus and the gain is high (Figure 5.4b). If, on the other hand, subjects perceive distinct causes, they seem to not only rely on the auditory stimulus but seem to get pushed away from the visual stimulus and exhibit negative gains. This is a very counterintuitive finding as previous models [Jacobs, 1999, van Beers et al., 1999, Ernst and Banks, 2002, Alais and Burr, 2004] predict only positive gains. Causal inference shows very similar behavior to the subjects' behavior; it also exhibits negative gains, and explains 99% of the variance of the gain. The causal inference model thus also predicts counterintuitive negative gains.

How can an optimal system ever exhibit negative gains? We consider the case where the visual stimulus is 5° to the left of the center and the auditory stimulus is in the center. On some trials, the lack of precision of the auditory system will lead to the perception of the auditory signal close to the visual signal, and then the system infers a common cause (Figure 5.4c red). On other trials, the auditory signal will be perceived far away from the visual signal, resulting in the model inferring distinct causes.

Due to this selection process, when a common cause is inferred, the perceived auditory location must be close to the visual stimulus, resulting in high gain. When distinct causes are inferred, the perceived auditory location must be far away from the visual stimulus, and the gain thus becomes negative. If instead of fitting the parameters, we assume that subjects use the true standard deviation of cue positions as σ_{pos} and $p_{common} = 0.5$, we still explain 85% of the variance of the gain with a parameter free model. Thus, the model is robust to changes in the parameter values.

In addition to the average responses of subjects, the standard deviation of responses can be informative about the strategy used by human subjects to combine cues. The standard deviation of human responses (Figure 5.4d), i.e., how imprecise the responses are, is also a function of the spatial disparity between the two stimuli. The model explains 82% of this variance and also explains the nonlinear behavior of the standard deviation.

5.5 Discussion

We have shown that the causal inference model predicts a wide range of effects observed in cue combination experiments many of which are unaccounted for by previous models [Jacobs, 1999, van Beers et al., 1999, Ernst and Banks, 2002, Alais and Burr, 2004] because

they predict constant weights for the combination of visual and auditory information. Each of our datasets rejects this prediction and shows that the gain varies in a nonlinear way. These models thus cannot predict any of the disparity-dependent changes of gain that we present here.

The model presented here can be seen as an elaboration of previous models that modeled the interactions between modalities by cross-modal interaction as a Bayesian model with a full 2-dimensional prior [Shams et al., 2005, Ernst, 2006, Roach et al., 2006], as in Chapter 3. The model we presented can also be rewritten as such an interaction model (see Section C.3). However, it dramatically reduces the number of free parameters because the prior now directly derives from the causal structure. But most importantly we can make predictions about the causal structure with the new model presented here that would have been impossible with the previous models [Jacobs, 1999, van Beers et al., 1999, Ernst and Banks, 2002, Alais and Burr, 2004, Shams et al., 2005, Ernst, 2006, Roach et al., 2006].

As the causal inference model uses a single inference rule to account for the entire spectrum of sensory cue combinations, many previous models are special cases of the model presented here, including those showing statistical optimality of full cue combination ($p_{common} = 1$) when the discrepancy between the signals is small [Jacobs, 1999, van Beers et al., 1999, Ernst and Banks, 2002, Alais and Burr, 2004]. In that case, $P(common|cues)$ will be very large.

Previous studies have shown other cases where the classical model of cue combination breaks down. E.g. as mentioned in Chapter 3, in the sound-induced flash illusion the number of perceived beeps is not generally identical to the number of perceived flashes [Shams et al., 2005]. Also earlier studies examined the breakdown of the ventriloquist effect [Howard and Templeton, 1966], reporting a loss of integration with increasing time delays [Wallace et al., 2004, Lewald et al., 2001]. In both of these effects the visual estimate is not identical to the auditory one and thus disagrees with the traditional model [Jacobs, 1999, van Beers et al., 1999, Ernst and Banks, 2002, Alais and Burr, 2004]. Various studies have started investigating the neural basis of cue combination using theoretical [Deneve et al., 2001] or experimental [Andersen, 1997, Wallace et al., 1992, 1998, Perrault Jr et al., 2003] methods.

Problems that are analogous to the problem of causal inference for cue combination also occur in the within-modality binding-problem whereby the nervous system has to determine

which set of stimuli correspond to the same object and should be bound together [Treisman, 1996, Reynolds and Desimone, 1999, Knill, 2003]. Our model may be seen as a partial solution to optimal binding. We are usually surrounded by many sights, sounds, odors, and tactile stimuli, and we constantly estimate if signals have the same cause.

In the study of higher-level cognition, many experiments have shown that people, starting from infancy, interpret events in terms of the actions of hidden causes [Buehner et al., 2003, Gopnik et al., 2004, Saxe et al., 2005, Griffiths and Tenenbaum, 2005, Waldmann, 2000]. If we see a window shatter, something or someone must have broken it; if a ball flies up into the air, something launched it. It is particularly hard to resist positing invisible common causes to explain surprising conjunctions of events [Gopnik et al., 2004, Griffiths and Tenenbaum, 2005], such as the sudden occurrence of several cases of the same rare cancer in a small town. These causal inferences in higher-level cognition may seem quite different than the causal inferences in sensory integration we have studied here: more deliberate, consciously accessible, and knowledge-dependent, rather than automatic, instantaneous, and universal.

Yet an intriguing link is suggested by our finding that optimal statistical principles can explain causal inference in sensory integration, as very similar principles have recently been shown to explain more conventional hidden-cause inferences in higher-level cognition [Gopnik et al., 2004, Griffiths and Tenenbaum, 2005, 2007]. Problems of inferring common causes from observed conjunctions arise everywhere across perception and cognition, and the brain may have evolved similar or even common mechanisms for performing these inferences accurately, in order to build veridical models of the world's underlying structure.

5.6 Summary

We have here shown that the idea of causal inference can be used to study perceptual as well as cognitive processing in the brain. Assuming that the brain utilizes a causal structure allows us to more specifically find the structure of the prior that we were looking for in Chapter 3. In the following we will examine this relationship further and see how far we can take predictions from Bayesian modeling.

Chapter 6

Testing Causal Priors

In everyday life we often think of bias as something to avoid, as in being biased against someone. In most experiments testing models of human perception, bias has indeed been thought of as implying an imperfection in the human nervous system. However, when operating with limited information, utilizing bias can be shown to be optimal using a Bayesian framework. When studying human psychophysical performance, the Bayesian framework allows the experimenter to test whether this bias is being utilized optimally. In Chapter 5 we have shown that a well-known audio-visual illusion, the ventriloquist illusion, can be described by a specific Bayesian model derived from causal inference principles that specifies a prior that assumes that stimuli are either from one or two causes. A requirement of any prior distribution is that, as the name implies, it is prior to any stimuli, and therefore not dependent on the stimulus. We are here the first to confirm that the estimated prior probabilities are indeed independent of the evidence (likelihood), which further indicates that human auditory-visual perception is a Bayesian inference process.

6.1 Introduction

Walking on a dark night in a heavy fog, you hear a honking sound and see a dark shape ahead. Is it a car speeding towards you or is the dark shape unrelated to the sound? This example highlights a basic challenge regarding multi-sensory processing. Our senses are seldom stimulated one at a time; multi-sensory input is typical, whether it is of people talking or cars passing in the street. Given information from multiple modalities, the brain needs to determine which signals should be considered as arising from one source and therefore need to be integrated, and which come from different sources and should be

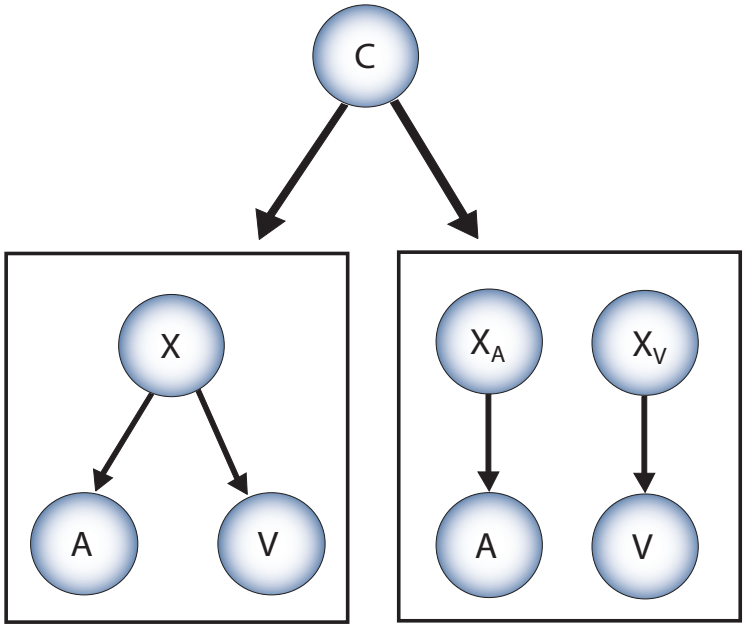


Figure 6.1: Graphical model describing the ideal observer. This is an alternative representation of the same model as presented in Chapter 5.

segregated.

Several previous studies have been able to provide a probabilistic account [Bulthoff and Mallot, 1988, Knill, 2003, Yuille and Bulthoff, 1996] for subject performance on a range of multi-sensory tasks including visual-haptic [Ernst and Banks, 2002], sensory-motor [Ghahramani et al., 1988], and visual-proprioceptive [van Beers et al., 1999] tasks. However the majority of these studies assumed that subjects combined the stimuli in an unbiased way, i.e., with a non-informative (flat uniform) prior. In other studies, the prior has been estimated, but there has been no way of comparing it across conditions [Stocker and Simoncelli, 2006].

Here we again use the ventriloquist illusion, to further examine how human observers use Bayesian inference to optimally combine prior information and sensory stimuli. We presented observers with auditory and visual stimuli presented at variable locations and asked them to report their perceived visual and auditory locations. We then compared these responses to an ideal-observer model derived from Bayes theorem and compared the fitted prior distributions across experimental conditions. Part of the experimental data that we analyze here was already presented in the previous chapter.

6.2 Methods

6.2.1 Stimuli

Experimental methods were similar to the last chapter. Visual and auditory stimuli were presented independently at one of five positions. The five locations extended from 10° to the left of the fixation point to 10° to the right of the fixation point along a horizontal line 5° below the fixation point, at 5° intervals. Visual stimuli were 35 ms presentations of Gabor wavelets extending 2° on a background of visual noise. The visual contrast was adjusted on an individual basis so that subjects unimodal performance was 90% correct for the high-contrast data (experiment 1) and 50% correct for the low-contrast data (Exp. 2). Auditory stimuli, synchronized with the visual stimuli in the auditory-visual conditions, were presented through a pair of headphones (Sennheiser HD280) and consisted of 35 ms white noise. The sound stimuli were filtered through a Head Related Transfer Function (HRTF), gathered individually from subjects using a pair of in-ear microphones (Sound Professionals) using procedures similar to those described by http://sound.media.mit.edu/KEMAR.html (again, see Section C.4), and simulated sounds originating from the five spatial locations in the frontoparallel plane where the visual stimuli were presented.

6.2.2 Procedure

In each trial, observers were asked to report both the perceived visual and auditory positions using the keyboard. Auditory and visual stimuli were presented alone or simultaneously, leading to a total of 35 conditions. The experiment consisted of 15 trials of each condition, amounting to a total of 525 trials, ordered pseudo-randomly. Twenty naive observers (undergraduate students at Caltech, eleven male) participated in the experiment. Observers were seated at a viewing distance of 54 cm from a 21-inch monitor. A fixation cross was always present 5 cm above the level of stimuli, but its color turned from red to white 0.5 second before the presentation of the stimuli, and then remained that color during the presentation. Participants were encouraged to take breaks every 10 minutes. Subjects were presented with the high-contrast data during one session, and then came back one week later, when they were presented with the low-contrast data.

6.2.3 Estimation of the likelihoods and priors

The causal inference model, as presented in the last chapter, requires combining the likelihood functions, $P(A|X_A)$ and $P(V|X_V)$, and the priors $P(X_A, X_V)$. In accordance with causal inference we assume a specific shape of the prior, requiring us to only estimate two parameters, the prior probability of there being a single cause, P_{common} and the variance of the prior, σ^2_{prior}. We here assume that the likelihoods and prior are all normally distributed with means μ_A, μ_V, and μ_{prior} and variances σ^2_A, σ^2_V, and σ^2_{prior} respectively. We furthermore assume that the mean of subjects likelihoods are at the veridical locations and the mean of the prior is at 0°, due to symmetry. Similar to Chapter 5, but in contrast to Chapter 3, we assume that subjects try to limit their mean deviation and therefore report the mean of their posterior. The 4 parameters (p_{common}, σ^2_A, σ^2_V, and σ^2_{prior}) were fitted to the subjects' responses using a 5000 trial Monte Carlo simulation and MATLABs fminsearch function (Mathworks, 2006). Implicitly here we make the natural assumption that the variability in the subjects' responses is due to the likelihood representing the subjects' uncertainty about the stimuli.

In this formulation it is difficult to directly see the shape of the prior, however it is possible to transform the current formulation into one where we more explicitly can show the shape of the prior as being a sum of a diagonal and normal distribution with no covariance (see Section C.3).

6.3 Results

In each trial, observers were presented with either a brief visual stimulus in one of five locations along the horizontal direction in the frontoparallel plane, or a brief sound at one of the same five locations, or both simultaneously. Their task was to report the location of the visual stimulus as well as the location of the sound in each trial. All 35 combinations of these stimuli (except for no flash and no sound) were presented in pseudo-random order.

6.3.1 Behavioral

Figure 6.2 shows that only in conditions in which the visual and auditory stimuli are identical (i.e., the conditions displayed along the diagonal) do observers consistently indicate perceiving the same location of events in both modalities. In conditions in which the incon-

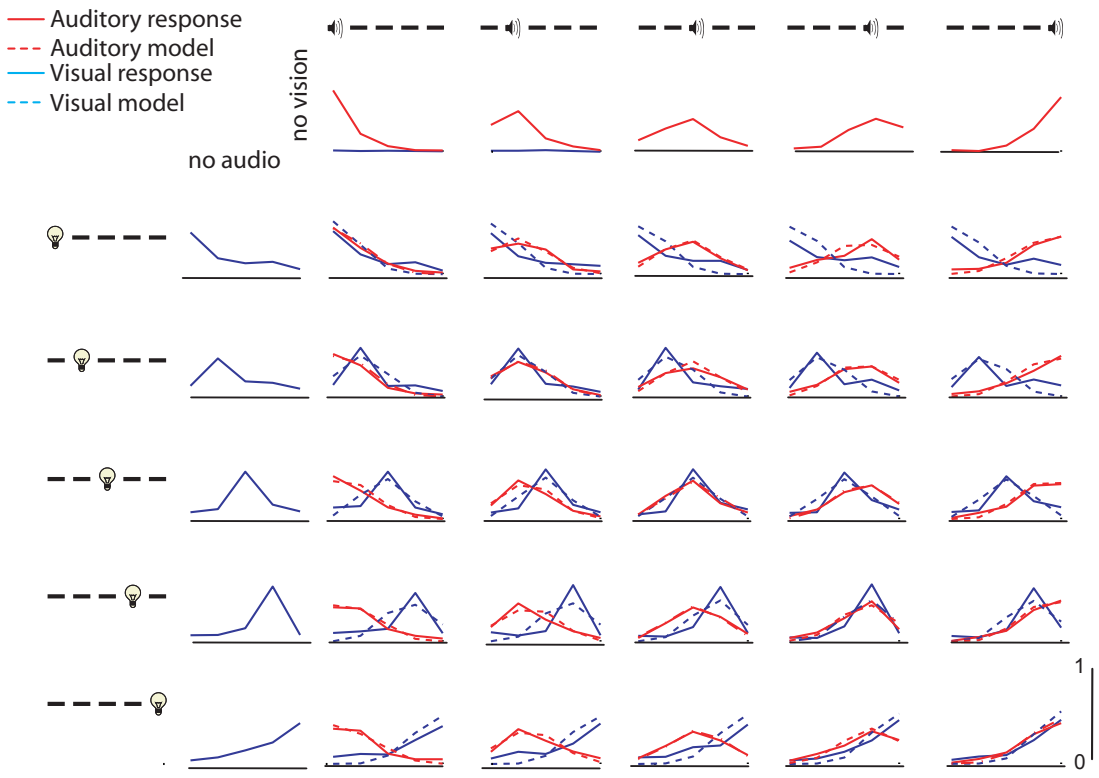


Figure 6.2: The average human observer responses (solid lines) are shown along with the predictions of the ideal observer (broken lines) for each of the 35 stimulus conditions. These plots show how often on average each button was pressed in each of the conditions.

sistency between the auditory and visual stimuli is not too large (e.g., in V1A2, visual and auditory stimuli in positions 1 and 2, respectively or V3A1 conditions), there is a tendency to combine the two modalities, as indicated by a shift towards the diagonal in Figure 6.2. In a previous experiment involving a temporal task (see Chapter 3 [Shams et al., 2005]), we found that the fusion of the auditory-visual percepts is achieved in these conditions of moderate discrepancy by a shift of the visual percept in the direction of the auditory percept, corresponding to the sound-induced flash illusion. Conversely, in the current experiment, the fusion is achieved by a shift of the auditory percept in the direction of the visual percept, corresponding to the ventriloquist illusion.

6.3.2 Bayesian ideal observer

Figure 6.1 shows the statistical structure of the causal inference model used, as presented in the previous chapter. The ideal-observer model does not assume integration a priori, but instead allows the nervous system to infer the causal structure from the input. Given stimuli A and V, the model tries to infer whether to believe in there being one or two causes of the stimuli. The responses are therefore a mixture of the one-cause and two-cause assumption, weighted by their inferred probability:

$$\hat{x} = P(common|V,A)\hat{x}_{common} + (1 - P(common|V,A))\hat{x}_{\neg common} \tag{6.1}$$

where $\hat{x}_{common}$ is the mean of the posterior $P(X|A,V)$ given that there is only one source and $\hat{x}_{\neg common}$ is the mean of the posterior $P(X|A,V)$ given that there are two sources.

The probability weights $P(common|V,A)$ are calculated in Appendix C.1.5; for more details I refer the reader to the methods section of the preceding chapter, or Sections C.1.2 and C.1.3. In order to compare the model with our experimental data, parameters for the likelihoods and prior were estimated by fitting the model to the data by minimizing the squared distance between the data and the model predictions.

To facilitate comparison of human and ideal-observer data, in Figure 6.2 only the two one-dimensional projections of the joint posteriors are displayed. The human observers performance is consistent with the ideal observer in all of the conditions, yielding a very small difference (sse =1.49,r^2=0.84), accounting for 300 data points (12 response combinations at 25 bimodal conditions) with 4 parameters. However this performance is far from being

as good as the one for the high-contrast data set, partly due to more variance in this data set.

A natural competitor is a model that assumes no interaction, $p_{common} = 0$. As expected, this led to a lower average performance (sse =2.95, $r^2 = 0.76$), further indicating that the priors deviate from uniform for this task.

For comparison we also fitted Gaussian distributions directly to each response distribution for each bimodal condition, requiring 2x2x25 parameters This led to a mean-squared error of 2.48, a worse performance than our model, despite using 100 parameters.

6.3.3 Independent encoding of priors and likelihoods

The results presented above indicate that the auditory-visual percepts of human observers are consistent with those of a Bayesian ideal observer. However, these results do not establish how the nervous system arrives at these percepts, whether indeed the percepts are produced by combining the likelihoods and priors, as in Bayesian inference, or whether the nervous system has an alternative algorithm for arriving at these optimal percepts.

We thus tested a central feature of Bayesian inference, the independence of the likelihoods and priors. The likelihoods $P(A|X_A)$ and $P(V|X_V)$ are functions of the input, whereas the prior probabilities $P(X_A, X_V)$ are assumed to be independent of the stimuli [1]. Changing the likelihood functions should have no effect on the priors, and, likewise, the same priors should be applicable to any stimulus. Whereas previous models did not have any way to test this, it is straightforward to examine this issue with our paradigm. Including the data from the last chapter we have two data sets acquired from the same group of subjects, for different visual contrasts, recorded with a weeks difference.

The Bayesian model predicts a change in the likelihood functions of the observers, which would in turn be reflected in a change in posteriors, but no change in the priors. Considering that the priors are estimated from the subjects responses in our model, it is not to be expected necessarily that a dramatic change in posteriors would leave the priors unchanged. Therefore, such a finding would provide strong support for the Bayesian proposition that the nervous system makes inference by combining likelihoods and priors.

As expected, the change in visual stimulus contrast led to considerable change in the

[1] As discussed previously and in Section C.3 we can represent the prior in the causal inference model as a sum $P(X_A, X_V) = p_{common} P(X = X_A = X_V) + (1 - p_{common}) P(X_A) P(X_V)$

visual performance, subjects on average decreased their visual performance in the visual-alone conditions by 41 percent. As mentioned optimizing the model directly on the low-contrast data set gives a good performance, although lower than for the high-contrast data, due to the larger variance in the visual data (see Table 6.1). Looking at the parameters of the optimal set, it is striking how similar the set for the high-contrast data is to the parameter set for the low-contrast data. Only the visual standard deviation changes more than 0.1. This is perfectly in accordance with the prediction of the theory that only the visual parameter should be modified.

To test this further we performed the analysis on a single subject basis, and again found that subjects' priors are very consistent. A 2-sided t-test[2] indicated that the only parameter to significantly change was the visual standard deviation ($p< 0.0005$), in accordance with the result for the mean data set and the theory.

Table 6.1: Parameters and results for the low- and high-contrast ventriloquist experiment, optimized on all data. For the individual fits, we indicate the mean ±standard error.

	SSE	P_{common}	σ_V	σ_A	σ_P
High contrast (all)	0.541	0.226	1.57°	7.15°	10.8°
Low contrast (all)	1.49	0.200	4.98°	7.00°	11.0°
High contrast (indiv)	2.83 ± 0.19	0.262 ± 0.05	$1.35 \pm 0.2°$	$8.85 \pm 1.0°$	$10.5 \pm 2.3°$
Low contrast (indiv)	4.66 ± 0.31	0.238 ± 0.05	$8.80 \pm 1.8°$	$9.70 \pm 1.95°$	$14.8 \pm 2.0°$

6.4 Discussion

In this chapter we find that a Bayesian model utilizing causal inference, is again able to account well for subjects' responses in a spatial audio-visual task. Further, when comparing the optimal parameter set for the high-contrast data set versus the low-contrast data set it is obvious that the only parameter to change is the variance of the visual likelihood. This implies that subjects use the same prior on the two data sets and a statistical group analysis confirms this finding, leading us to conclude that the priors are indeed independent of the stimuli and therefore true priors.

Our finding that observers priors are independent of the stimuli is the first confirmation of this central prediction of all Bayesian models and provides further support for the hypothesis that the brain uses a mechanism that follows Bayesian inference for combining

[2]The p-value was corrected for restricting the values to be positive by numerical simulation.

auditory and visual signals. A change in the visual stimulus, which should lead to a change in the likelihood functions, and was manifested by a significant change in the visual performance, left the estimated priors unchanged. This indicates that the priors and likelihoods are represented separately in the nervous system as predicted by the Bayesian model and are combined according to Bayes' rule in this perceptual task.

As our model uses a single inference rule to account for the entire spectrum of sensory cue combination, many previous models are special cases of the model presented here. The present model is also consistent with previous findings showing the need for non-uniform priors in the ventriloquist illusion [Battaglia et al., 2003], and those showing statistical optimality of multi-sensory integration when the discrepancy between the signals is small [Ernst and Banks, 2002, Ghahramani et al., 1988, van Beers et al., 1999] - the latter being a special case of the present model.

One question we are not addressing here is how the priors are encoded in the brain - whether hard-wired at the synaptic level, or due to firing patterns of specific groups of neurons. Related is also the question of whether it is possible to modify a perceptual prior.

A possible future direction for this research is to investigate how modulations of multi-sensory integration can be caused by changes in attention or relative timing of the signals. It is known, for example, that as the time difference between auditory and visual stimuli increases, integration gracefully degrades [Shams et al., 2002]. To date, none of the models of cue combination have addressed these factors, which almost invariably affect cue combination in all tasks. Incorporation of these factors may require modeling how Bayesian inference is achieved in the nervous system at a neuronal level. Plausible candidates for a neural implementation of two-dimensional Bayesian inference with priors include basis function networks with attractor dynamics [Deneve et al., 2001, Pouget et al., 2002] and Helmholtz machines [Abbott and Dayan, 2005].

6.5 Summary

We have here shown that one of the major predictions from Bayesian inference, the independence of likelihood and priors is indeed valid for this task. This is strong support for the thesis that subjects perform in accordance with the Bayesian model.

However we have only examined audio-visual, cross-modal processing so far, whether the results extends to within-modal processing will be the subject of the next chapter.

Chapter 7

Within- versus Cross-Modal Processing

So far we have seen that a Bayesian and causal inference models are able to explain subjects' performance on two audio-visual tasks. Here we return to the audio-visual flash illusion to examine how the causal model performs on this data set. This allows us to compare cross-modal to within-modal processing by performing a visual-visual variant of the flash illusion and to examine the importance of the prior.

7.1 Introduction

A number of studies have examined how human subjects can combine information across [Battaglia et al., 2003, Ernst and Banks, 2002] and within modalities [Mamassian and Landy, 2001, Jacobs, 1999], a large number of them finding that subjects perform very close to optimally, as derived from Bayes' theorem.[1] The majority of these have tended to only study situations where two cues to the same property is presented with small discrepancy, to facilitate the fusion of the two. However, under natural conditions being presented with several statistically independent cues is common, for which situation fusion would be detrimental.

As a result, a more recent interest has arisen for studies where subjects are tested for cue interactions. By presenting subjects with independent potentially conflicting signals, and asking subjects to segregate them, researchers can study to what degree such interactions can be explained by a Bayesian model [Shams et al., 2005]. We have previously presented such a model and shown that it is able to explain subjects' behavior on an audio-visual flash

[1] Although not all. For a counter example see e.g., [Hammett et al., 2007].

illusion task where subjects were asked to report both their visual and auditory percept (see Chapter 3). Such a model can be shown to be a super-set of the previous forced fusion models, and can therefore also account for the results from studies using such models.

The first question we wish to answer is whether we can explain the data from the audio-visual flash illusion with the more restricted causal inference model. Secondarily we wish to compare these results with the results of a visual-visual experiment in the hopes of understanding the difference between within- and cross-modal processing.

Whether within- and cross-modal stimuli interacts in the same way, is a topic under discussion [Hillis et al., 2002, Gepshtein et al., 2005]. Results suggest that within-modal processing should not be considered in a category completely separate from cross-modal processing, but that there may be a more gradual difference.

7.2 The auditory-visual flash illusion revisited

An immediate question before doing any type of comparison, is how the causal inference model performs on the previous data sets for the audio-visual flash illusion. We therefore decided to test the causal inference model on the data. Our expectation as to the result is not unequivocal. The prior in Chapter 3 was unrestricted, but was estimated, not fitted. Furthermore the more stringent use of the utility function and the calculations that it requires in the causal inference model will influence the results.

We applied the model, similarly to Chapter 5, with a few exceptions. In the ventriloquist experiment we felt justified in requiring the prior to be centered at 0 degrees, relative to the observers gaze, as we would otherwise have left/right bias. In the flash experiment we have no similar symmetry considerations to allow us to fix the mean of the prior, which was therefore fitted as well. We therefore fit 5 parameters p_{common}, σ_{pos}, μ_{prior}, σ_{vis}, and σ_{aud} by minimizing the squared difference between the model predictions and subjects' responses (sse).

The result of the fit is a summed squared error of 1.08 - worse than the result from the Bayesian model with full 2D prior (sse =0.804), that we used in Chapter 3. Parameters are shown in Table 7.1 below.

As the causal inference model is a nested model of the full 2D prior model, we were prepared for the possibility of a lower performance, and, judging from this alone, the causal

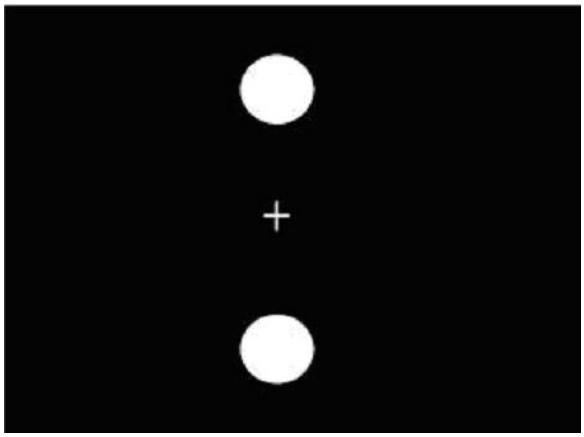


Figure 7.1: Example of presentation screen. A visual circle would appear for 10 ms, 12 degrees above and/or below the fixation point.

model would seem to be inferior. However, when we return to the subject of model comparison in Chapter 8, we will see that things are a bit more complex.

However as the current focus is to examine the difference between within- and cross-modal processing, we are now ready to examine the within-modal part.

7.3 A visual-visual flash illusion

The studies so far have all been based on two types of audio-visual illusions. However the flash illusion can very easily be converted into a visual-visual illusion, making comparison easy (see also [Bowen et al., 1987]). We explain the experiment and present the results of performing the fitting of the causal model.

7.3.1 Experimental methods

Thirteen naive observers (undergraduate students at Caltech, 8 male) participated in the experiment. Observers were seated at a viewing distance of 54 cm from a 21-inch monitor. White circles of 1.5 degrees diameter were presented 12 degrees above and below a fixation point and flashed independently 1 to 4 times for 10 ms with a 70 ms interval (see Fig. 7.1). In each trial, observers were asked to report the perceived number of flashes both above and below the fixation point. Stimuli were created and subjects' responses were recorded using MATLAB (Mathworks, Mass.) and the Psychophysics Toolbox [Brainard, 1997].

7.3.2 The ideal observer model

To better analyze the data we used our previous causal inference model from Chapter 5. The model is a mixture model, combining a sub-model that assumes a single source, and a sub-model that assumes two separate sources. For each stimulus set the model estimates the likelihood of the stimuli originating from a common source, and given this, calculates the number of flashes (X_1, X_2). We assume that subjects try to minimize their mean squared error and that they will therefore use the mean estimation of their distributions. Again we assume that we can use normal distributions to describe the likelihoods and prior. In order to establish the model, we need to fit 4 parameters for the model, the prior probability of the stimuli originating from a common source p_{common}, the prior mean and variance of the number of flashes, μ_{prior}, σ^2_{prior}, and the variance of the visual likelihood of the upper stimuli/lower stimuli σ^2_{vis}.

7.3.3 Equations

The modeling approach is very similar to what we have seen in the last two chapters, the largest difference being that we now have no auditory likelihood, but instead have 2 visual stimuli, V_1 and V_2. The model estimates the probability that the stimuli originated from a common source using Bayes' theorem:

$$P(C|V_1, V_2) = \frac{P(V_1, V_2|C)P(C)}{P(V_1, V_2)} \tag{7.1}$$

where C is short for common. For calculation of these variables see Section C.1.5. To estimate the number of flashes, the model weighs the probability of each sub-model's estimate.

$$\hat{x} = P(common|V_1, V_2)\hat{x}_{common} + (1 - P(common|V_1, V_2))\hat{x}_{\neg common} \tag{7.2}$$

The two estimates $\hat{x}_{common}$ and $\hat{x}_{\neg common}$ are calculated from traditional methods (see equations (2.6) and (2.12)).

7.4 Results

Thirteen subjects participated in this experiment allowing us to compare their theoretical posterior and response patterns.

7.4.1 Experimental results

Figure 7.2 illustrates the result of the experiment by plotting the average response probabilities of subjects, given a stimulus set, V1 and V2. When the number of visual flashes are not too different, there is a strong tendency for subjects to combine the two stimuli into a single estimate, e.g., when presented with one flash above, two below (V1=1, V2=2 condition) subjects often report seeing 1 flash above as well as below. For conditions with larger discrepancy subjects have a much lower chance of combining the stimuli to a single response, but still have a much lower chance of picking the veridical number of flashes than when processing the stimuli independently.

In our previous similar audio-visual study (see above and Chapter 3) there was a clear difference between the two modalities, the auditory performance was much more precise than the visual. In the current study, as we are dealing with within-modal comparison, we expected the two responses to be very similar, as is indeed evident by inspecting Figure 7.2. Notice also that the shifts away from veridical are larger in this data set.

7.4.2 Model results

Figure 7.2 shows the model's results on top of the subjects' responses. The model is capable of replicating the subjects' performance well, with little difference between the curves in Figure 7.2 (sse=0.735) and explains a large amount of the variance in the 2D posterior, r^2=0.90. However it is off in a few places where it seems to underestimate the amount of interaction, e.g., condition (V1=1, V2=2). To get a better feel for the model's performance we tried two alternative models: a flat prior model and a Gaussian fit model. When we applied the model with a flat prior on the same data (only one free parameter, σ_V), we find a much lower performance ($sse = 2.70$). Clearly the prior in this case is very different from flat. As a very different alternative we fitted a pseudo-model: for each condition and each modality we fit a normal distribution with mean μ_i and variance σ_i^2. This model uses 16*2*2=64 parameters, so we would expect it to do reasonably well, compared to our

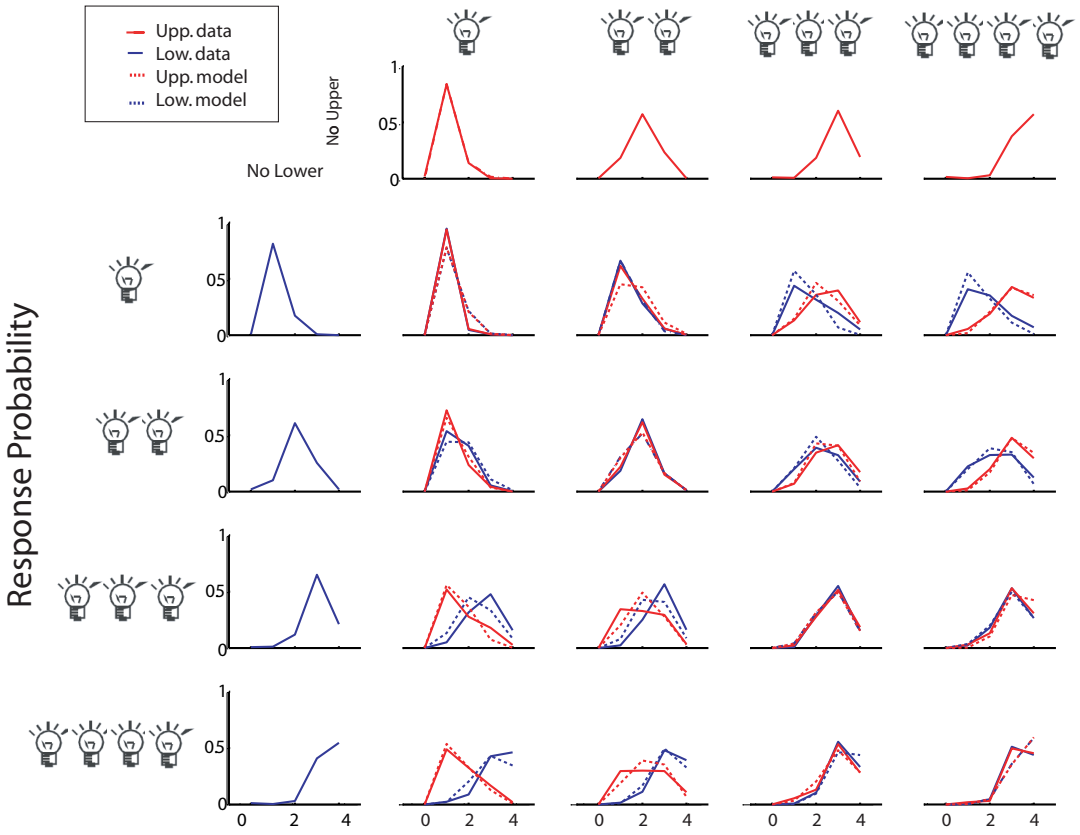


Figure 7.2: Comparison between subjects and model. (5 x 5 -1) conditions were presented to subjects, row conditions represent flashes above the fixation cross, columns those below the fixation (check). For each condition we plotted the probability of subjects responding 0 through 5 in complete lines, and the probability the model predicts them responding 0 through 5 as broken lines. Red is responses for the flashes below the fixation, blue is for above.

4-parameter causal inference model, however its performance was lower ($sse = 1.45$).

Overall the causal inference model does a very good job.

7.5 Within- versus cross-modal

We have performed two similar experiments, one where the interactions were within-modal (visual-visual flash illusion), and one where the interactions were cross-modal (auditory-visual flash illusion). Care was taken to keep the setup as similar as possible between the two experiments. As we have also analyzed them with the same model, the causal inference model, we can now compare them as in the table below.

Table 7.1: Parameters and results for the audio-visual and visual-visual flash illusion experiment using the causal inference model.

	SSE	p_{common}	σ_V	σ_A	σ_P	μ_P
Audio-Visual	1.075	0.455	0.815	0.451	0.983	-1.377
Visual-Visual	0.735	0.631	1.012		2.410	0.527

Both illusions are well described by the model - the visual-visual slightly better, considering it also has one less free parameter. This seems to indicate that both within- and cross-modal perception can be described by Bayesian modeling.

The difference in number of parameters is due to our assumptions that the likelihood functions for the visual signal above and below the fixation point is the same. An argument could be made for assigning different variances, due to evidence of different neural processing for upper and lower visual fields [Previc, 1990], however testing such a model provided very little improvement for the extra parameter ($sse = 0.717$).

It is interesting to see that the parameters do seem to change between the two experiments. Considering the experiments are similar - both tasks involve counting temporal occurrence - one might be tempted to expect similar parameter values. However the fitting does make a couple of simplifying assumptions that may not be valid across experiments, e.g., $P(X|common) = P(X|\neg common)$ or $P(X_V|\neg common) = P(X_A|\neg common)$. In the following chapter we will try relaxing such constraints.

One parameter where we do expect to see differences is the prior probability of experiencing a single cause, p_{common}. The task in these experiments are of a temporal nature, Synchronization of separate causes seem intuitively more likely to happen for visual-auditory

stimuli than for visual visual, implying that p_{common} should be higher for the visual-visual task, which is indeed what we observe.

Further more, this fit does not tell us whether the causal inference model with the restricted prior is better than the model with an unrestricted prior, as the difference in modeling could be influencing the results. In fact, since the causal inference model is a restriction on the full 2D prior, we would expect the 2D prior to be at least equally good, although less parsimonious. We therefore devised a way to directly compare different priors, as described in the next chapter.

7.6 Discussion

We have shown that a within-modal illusion, the double-flash illusion [Bowen et al., 1987] can be interpreted in terms of a full Bayesian causal inference model, with a prior very different from flat, to explain the strong interactions in the data. This means we are able to explain both within- and cross-modal illusions with such models.

The idea that we can explain differences between cross- and within-modal processing by the help of a single parameter, p_{common}, is intriguing and opens up a range of interesting extensions, e.g. can we predict what the value in an arbitrary experiment should be, based on first principles or statistics in the real world?

These results also seem to be in conflict with the results of Hillis et al. [Hillis et al., 2002], which found different processing for a within-modal (visual) and cross-modal task (visual-haptic). We will return to this in Chapter 9.

7.7 Summary

This chapter has shown a way of studying the difference between within- and cross-modal processing. Our data hint, if not prove, that the difference may lie in the strength of the prior $p_{commmon}$; i.e. that we can explain the stronger tendency for within-modal cues to combine as a higher prior expectation of a single cause.

This again shows the applicability of the causal inference model. However the model is a restriction of the full 2D prior model, and one can not help but wonder if it may be too limiting. We will examine this topic next.

Chapter 8

Model Comparisons

In the previous chapters we examined two Bayesian models and showed how they are able to describe human performance on several within-modal and cross-modal tasks. Here we will try to compare them, analytically and numerically.

8.1 Introduction

That a model is able to fit data well is certainly a necessary, but not sufficient condition for being a good model. Other requirements usually mentioned are simplicity, the ability to create new predictions, and the relation to the entire research framework or paradigm [Kuhn, 1962]. The ability of the models to fit the data has already been debated in the previous chapters; here I will address the problems of simplicity and relation to other models. The question of how a model fits into a more general framework/scientific study I will leave for the discussion in next chapter.

In Chapter 3 we saw that the Bayesian model with a full 2D prior is able to explain the data from the audio-visual flash illusion very well. Chapters 5, 6 and 7 showed that the causal inference model accounts extremely well for all the data sets and, for the ventriloquist data, that the parameters for the prior even carry across experimental conditions. How do these results compare and how do the two models relate?

Structurally we can show that the causal inference model is just a restriction on the full 2D prior model (see the appendix) - a nested model, if you will. However each model has used different utility functions and ways to generate the model predictions, and has been applied to different data sets, so a comparison is not easy. This chapter is trying to remedy that by formulating a parametric 'supermodel', that has a less restrictive parametric prior,

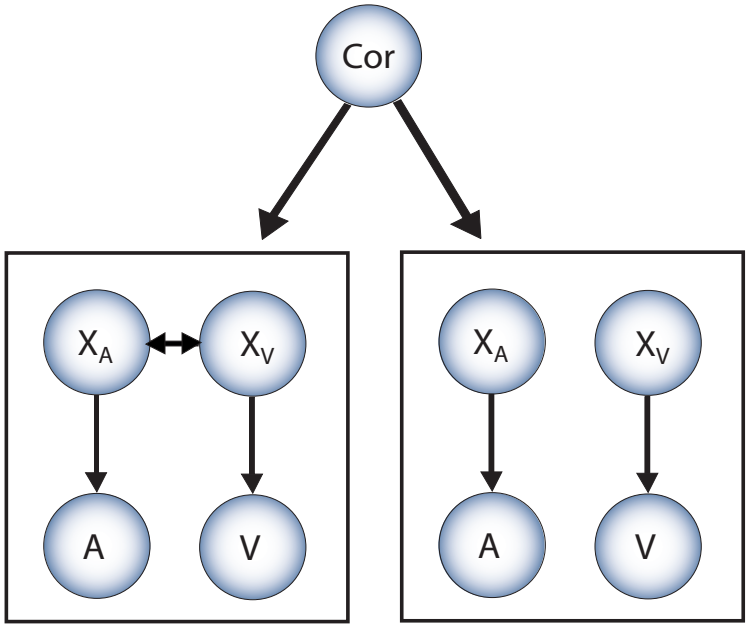


Figure 8.1: A graphical model representation of the full inference model. For every trial the positions of X_V, X_A, and the switching variable *correlated* are drawn randomly according to their distributions.

but uses the analytical framework and utility function from the causal inference model.

8.2 'Super' model

We devised a new more general model, as a parameterized alternative to the unconstrained model in Chapter 3. It has the advantage that several of the previous models, such as the fusion model (see Section 2.2) and the causal inference model (from Chapter 5), can be considered nested models within it.

In order to be able to create a prior distribution that is capable of taking as unconstrained a shape as possible, we make use of the Gaussian distribution as a radial basis function. Similar to the Fourier statement that any function can be replicated by a sum of sine or cosine functions, the radial basis theorem states that any arbitrary function can be replicated by a sum of Gaussian distributions with fitted mean and variance [Bishop, 1995].

To simulate the shape of a large range of priors, we therefore chose the sum of two Gaussian distributions, but restricted the covariance of the first Gaussian distribution to be 0. As Figure 8.2 shows, just two normal distributions are enough to recreate a large number

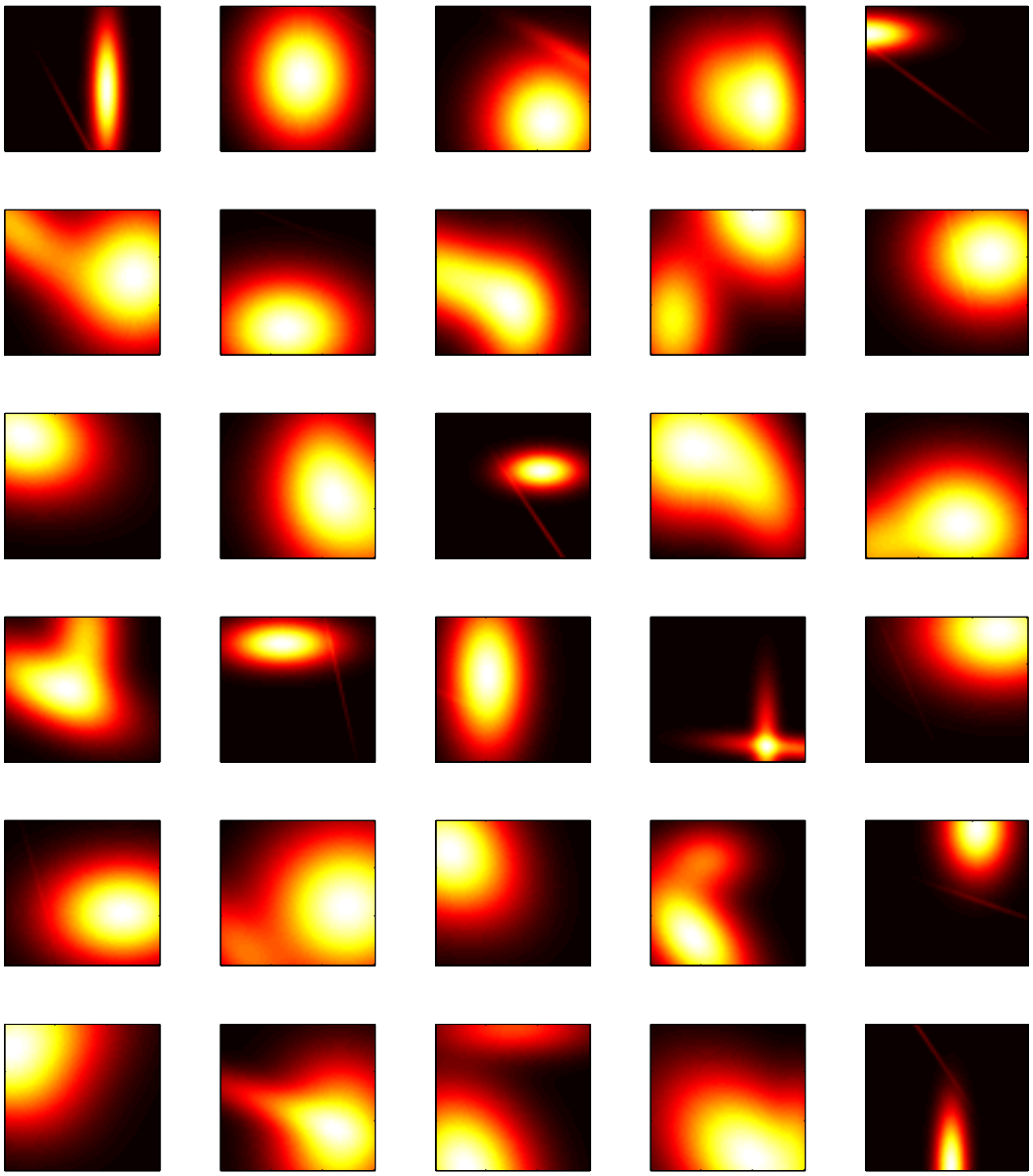


Figure 8.2: Examples of random priors possible by the addition of two Gaussian distribution, one with full covariance.

of functions. An added advantage is, of course, that the product of Gaussian distribution is also a Gaussian distribution, making our computations much simpler (see Appendix C.2.2).

The generative model can be thought of as follows: two causes are either uncorrelated (with probability $1 - p_{correlated}$), or share some correlation (with probability $p_{correlated}$). They each give rise to a stimulus, which the nervous system registers. The task of the brain is then to infer whether the two causes are likely to be correlated or not, and, given that, what the causes were. If we assumed full covariance in the correlated case (single cause), this would reduce to the causal inference model.

Similar to the causal inference model we have a mixture model:

$$\hat{x} = P(correlate|V, A)\hat{x}_{correlated} + (1 - P(correlated|V, A))\hat{x}_{\neg correlated}. \tag{8.1}$$

We assume that everything is normally distributed with likelihoods, $P(V, A|correlated)$ as a Gaussian distribution with full covariance matrix and $P(V, A|\neg correlated)$ as a Gaussian distribution with 0 covariance. The full version of this model has 12 parameters, $p_{correlated}$, σ_V, σ_A, $\mu_{\neg cor,V}$, $\sigma_{\neg cor,V}$, $\mu_{\neg cor,A}$, $\sigma_{\neg cor,A}$, $\mu_{cor,V}$, $\sigma_{cor,V}$, $\mu_{cor,A}$, $\sigma_{cor,A}$, and $\sigma_{cor,Covar}$.

If we restrict the mean of the Gaussian distributions to the center, set the covariance $\sigma^2_{Covar,cor}$ to be 100 percent, and set all the prior variances equal, $\sigma_{V,cor} = \sigma_{A,cor} = \sigma_{V,\neg cor} = \sigma_{A,\neg cor}$, we reduce this model to the version of the causal inference model we used in Chapter 5. Our number of parameters also decreases from 12 to 4.

The causal inference model therefore becomes a nested model within this 'Super-model.' If we further set $p_{correlated} = 1$ we have a variant of the forced fusion model. If instead we set $p_{correlated} = 0$), it is a model that assumes no interaction. Naturally the causal inference model is a mixture between these two.

For modeling purposes we restricted the number of parameters of this 'super-model' in the following way: We set the variances for the correlated Gaussian distribution equal, $\sigma_{cor,V} = \sigma_{cor,A}$. For the ventriloquist illusion data sets we set all the means to 0 degrees (straight ahead) due to symmetry; for the flash illusions we assumed the means would be equal. This reduced our number of parameters to 7 for the ventriloquist data sets and 8 and 6 respectively for the audio-visual flash illusion and the visual-visual flash illusion.

As the number of parameters increase, the optimization process also becomes more

Figure 8.3: Three different priors, corresponding to forced fusion, no interaction, and causal inference models.

complicated. We used 100 random seeds to the optimization process to ensure that the global minimum was found for each data set.

8.3 Results

We optimized such a model on each of our four previous data sets; the audio-visual flash illusion, the visual-visual flash illusion, and the high- and low-contrast data sets of the ventriloquist illusion. The resulting optimal parameters are printed in Table 8.1 and the corresponding priors are shown in Figure 8.4. As is clear from inspection, they all have a shape very similar to the one predicted by the causal inference model, providing support for the idea that it is both strong and parsimonious.

Table 8.1: Parameters and results for the 4 data sets when using a model that allows a larger number of parameters and flexibility of the prior distribution.

	SSE	p_{cor}	σ_V	σ_A	$\sigma_{\neg cor,V}$	$\sigma_{\neg cor,A}$	μ_{priors}	σ_{cor}	$Covar$
Ventrilo Hi	0.449	0.198	1.54°	6.85°	4.75°	9.50°		12.3°	100%
Ventrilo Lo	1.452	0.277	4.21°	7.07°	8.80°	11.5°		11.3°	86.1%
Audio-Visual	0.228	0.414	0.927	0.296	1.702	0.596	1.348	0.566	100%
Visual-Visual	0.725	0.627	1.009		2.782		0.040	2.457	100%

It is worth noticing here that the parameters for the two ventriloquist data sets start separating slightly at this point. Especially the standard deviation of the no-interaction Gaussian distribution is now different (4.75° vs 8.80°). This may just be due to over fitting, a danger that becomes more relevant as we add parameters.

Comparing the parameters across experiments it is also interesting to see that the prior probability of a correlated cause $p_{correlated}$ is quite different for the different illusions. It is hard to say anything conclusive based on merely this data, but as previously discussed in Chapter 7 this may be due to stronger binding for within- than cross-modal stimuli [Hillis et al., 2002]. We shall return again to this topic in the discussion.

Worth noticing also is that the covariance is very high - 100 percent for three of the data sets, 86 percent for the last. This seems to be strong support for the causal model idea, as this is one of the key requirements. Even when we relax the requirement that the covariance be full, it still finds this as the optimal solution.

Looking at the results for the audio-visual flash illusion it may seem like the full model is able to explain the data much better, the sum of squared error - *sse* goes from 1.08 to

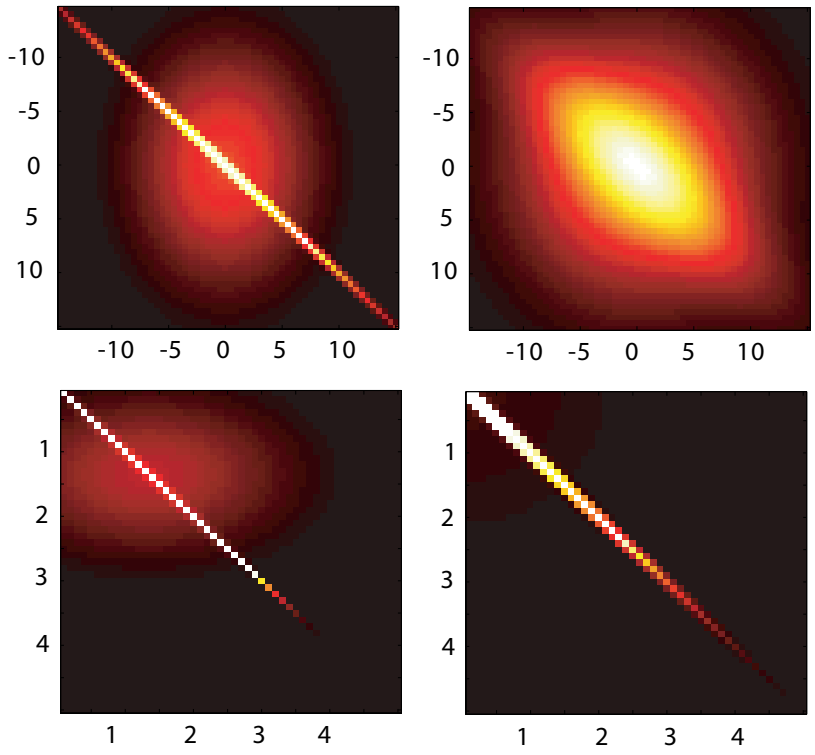


Figure 8.4: Priors optimized from the datasets. Upper: Ventriloquist high-contrast and low-contrast. Lower: Audio-visual and visual-visual flash illusion.

0.23, which is a huge change. However examining the model a little closer reveals that this improvement is mostly due to adding one more parameter by allowing the two variances of the independent Gaussian distribution, $\sigma_{\neg cor,V}$ and $\sigma_{\neg cor,A}$, to vary independently. Such a version of the causal model with 6 parameters, reduces the error to 0.257 for this data set. Reducing the covariance only tends to increase the error. That the variances are unequal is interesting, as it implies that the prior expected variability in numbers is larger for visual than auditory signals.

8.4 Discussion

For comparison, the performance of each model on each data set is summarized in Table 8.2.

Table 8.2: Results for our four data sets using different models. The number in parenthesis indicates the number of parameters fitted.

SSE	Flat prior	Causal model	'Super' model	Gaus fit
Ventriloquist Hi	1.25(2)	0.54(4)	0.449(7)	1.22(100)
Ventriloquist Lo	1.96(2)	1.49(4)	1.452(7)	2.48(100)
Audio-Visual	3.96(2)	1.0751(5)	0.228(8)	1.65(64)
Visual-Visual	2.70(2)	0.7354(4)	0.725(6)	1.45(64)

As we are using nested models, we can test whether the model with larger number of parameters is significantly better or is just noise. Using both a F-test and log-likelihood ratio test we find that though the 'super' model only reduces the error slightly (see Table 8.2), it is significantly better ($P < 0.05$) for the 2 ventriloquist illusions and the auditory-visual flash illusion, but not for the visual-visual flash illusion .

Clearly the 'Super' model has improved the fit, but that the improvement is moderate and that the covariance is full or close to full, is more support for the causal inference model.

8.5 Summary

The preceding results show that the causal inference model is not just supported by theoretical considerations, but also by comparison to other models. Although a better fit can be achieved with more parameters, the gain is relatively small, implying that the causal inference model is a good compromise.

Next we try to figure out what all this means.

Chapter 9

Discussion

Naturally, many questions remain at this point, as the application of Bayesian inference to problems of the nervous system is only gaining more traction in recent years. Is there a difference between within- and cross-modal processing? How does Bayesian inference apply to learning? Can we find direct proof of Bayesian computations being performed in the human brain? How would we expect such computations to be performed at a neural network level? And is the human brain really 'optimal'?

9.1 Introduction

In the previous chapters the reader has hopefully been convinced that Bayes' theorem has a use in studying perceptual processing, as described in theory and practice.

We demonstrated that rules derived from optimality principles can be used to describe subjects' behavior on several audio-visual tasks for very different conditions. The model can describe within- as well as cross-modal data and we showed that the prior beliefs are constant across conditions. We compared two models, one that performs causal inference and one with fewer assumptions about the structure of the prior, and showed that though the model with more parameters can explain more of the variance, the improvement is generally small.

9.2 Extensions of our results

Any answer tends to bring a range of new questions; this thesis is no exception. What further uses could this work have?, what studies would follow from this?, what are the

implications? The following is merely intended as a sampling of ideas.

9.2.1 Within- versus cross-modal

Although we have here touched upon the idea of comparing cross-model and within-modal processing, the difference is certainly not yet clear. The results presented here (in Chapter 7) suggest that the difference is very gradual, merely a difference in the prior probability of a single cause p_{common}.

Other studies have instead claimed an all or none approach - forced fusion or not. Hillis et al. [Hillis et al., 2002] presented subjects with either a visual and a haptic cue or two different visual cues, and asked them to judge the width of an object in a two-alternative forced-choice task (2AFC). In some trials the cues were consistent, and combining them would therefore be advantageous (leading to a better performance), in others the cues were inconsistent, and combinations would therefore be disadvantageous (leading to a worse performance).

They found that subjects' performance on the visual-visual task did indeed get worse as the two cues became more inconsistent, whereas subjects' performance on the visual-haptic task showed very little change. They chose to interpret this as the visual cues being forced to be fused, whereas the visual-haptic cue combination did not have this strong tendency to be fused together. An inspection of their data shows that this is clearly a simplification - for the consistent stimuli subjects may combine them, but for the inconsistent stimuli they clearly do not, although their performance does decrease.

In terms of our causal model, it is easy to interpret these results. According to this interpretation, the visual-haptic prior probability of a single cause $p_{common,VH}$ is smaller than the similar parameter for the visual-visual stimuli $p_{common,VV}$. Subjects therefore tend to combine their estimates more strongly, leading to a larger error. This is similar to the results from our VV and AV experiments for the flash illusion: a larger p_{common} for the visual-visual task than for the audio-visual.

It should also be emphasized that there is no other model that would be able to explain this data set.

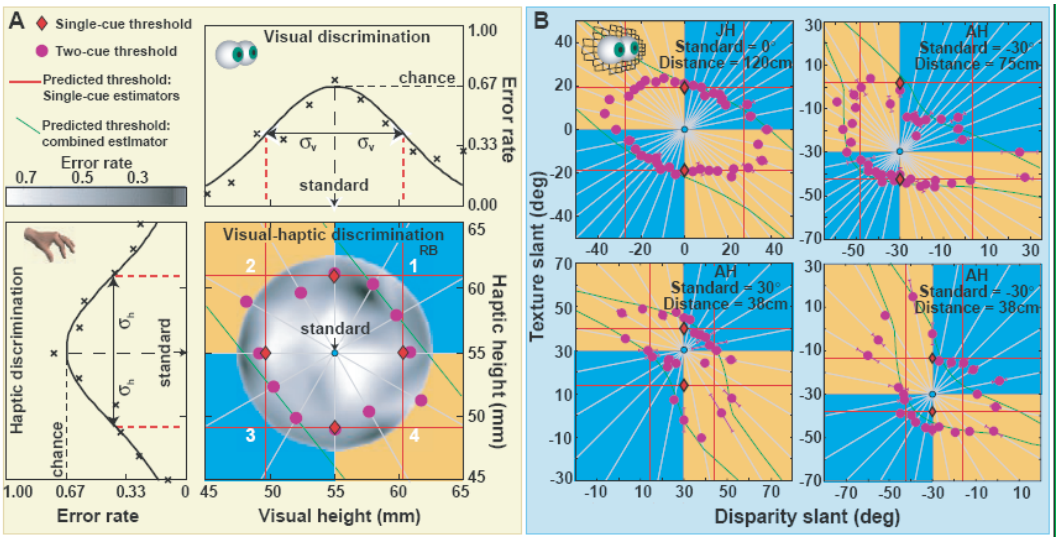


Figure 9.1: (From [Hillis et al., 2002].) Subjects were presented with visual and haptic cue combinations. a) Subjects judged visual and haptic cues, the standard deviation of their judgments falling close to the unimodal predictions (represented by the red lines), as predicted by a no-combination rule. b) For the visual-visual cue, the standard deviation of subjects' judgments would go outside the square (implying lower performance) in the incongruent conditions (orange quadrants), but would move inside (implying improved performance) for the congruent conditions (blue quadrants). Four examples are shown. For more details, see [Hillis et al., 2002].

9.2.2 Binding problem

Rather surprisingly, our results can also apply to the classical binding problem. The perceptual version of the binding problem asks how we connect events from different cue or separate modalities into one conscious percept. A typical example is of a baseball bat hitting a ball, causing us not to perceive the sound and the visual bouncing as distinct events, but instead unify them into one percept (for a review of the problem, see [Roskies, 1999]). How does the human brain 'bind' features together to create the percept of an object, as opposed to a string of single perceptions?

Causal inference gives a partial answer in that the causal structure gets inferred from the perceptual stimuli and prior expectations, and directly indicates if the properties are believed to be from a single cause. In the baseball example the brain would try to infer the causal structure, i.e., a single event versus two events, by combining the perception with prior expectations of spatial-temporal relations in the world (e.g., single events tend to cause audio and visual signals to be correlated in space, separate events do not). Therefore causal inference can solve the problem of how to decide if two properties should be combined, but does not answer how the actual unified 'feeling' of a single object is achieved. One theory currently involves synchronized firing in distinct brain areas [Singer, 1999], but the field is still wide open.

9.2.3 Prior

As we have seen previously we can show that the prior probabilities are constant across conditions as would be expected from Bayesian theory. This is another strong indicator of the generalizability of these results. However, no experiments to date have examined how to modify subjects' priors, in fact, very few studies have been able to estimate subjects' priors [Stocker and Simoncelli, 2006, Mamassian and Landy, 2001], other studies have assumed a specific shape, usually based on computational simplicity [Weiss et al., 2002]. Whether it is possible to modify a subjects' priors is therefore still an open question. If modifiable, in theory one would expect that prolonged exposure to particular stimulus distributions would be able to influence it, as the nervous system would learn the new distribution. However there were no indicators of any learning/adaptation in any of the experiments here presented.

We are of course being bombarded by audio-visual stimuli at every moment we are awake, so any new statistical distribution would have to counteract this, and there seems to be no good reason why it would be advantageous to modify the priors over such a short time period (say 30 minutes). Whether the priors are therefore encoded synaptically, or caused by some firing pattern is an open question [Ma et al., 2006]. A related question is in what frame of reference a prior is represented. For all the experiments conducted here, the eye-centered frame of reference and the head-centered frame of reference were aligned, as subjects would fixate straight ahead. How would the priors be influenced if we shifted the two frames, e.g., by simply asking subjects to fixate in the periphery? Can this tell us something about which neuronal pathway processes this input? Hopefully further experiments can answer these questions.

9.2.4 Bayes and learning

A natural extension of the Bayesian inference is for learning, since the posterior $P(X|V)$ is the new estimate of X after gaining new knowledge about the observed variable V. It is therefore not surprising that Bayesian learning has gained a lot of traction in the Machine Learning community, where it is used extensively. The most well known (and annoying) use has probably been for the Microsoft Office Assistant, also known as 'Clippy', but it is also used for a range of other purposes, including email filtering, and will undoubtedly be used even more in the future.

But Bayesian modeling has also been used to study human learning (see e.g., [Eckstein et al., 2004, Kording and Wolpert, 2004, Tenenbaum, 1999, Orbán et al., 2006]) and we will briefly touch upon the ideas of using Bayesian inference to study learning in Appendix A.

9.3 High-level and low-level

The potential link between the processing on a perceptual level and cognitive level, as hinted at by the causal inference model, opens up some interesting ideas. Ever since Kahneman and Tversky did their Nobel-prize-winning studies in the seventies [Tversky and Kahneman, 1974], human cognitive processing has been thought of as inefficient. They presented subjects with a range of tasks involving decision making based on uncertainty and showed that subjects tended to perform them based on some heuristic - that is, based on some simpler

algorithm that tended to make too simple assumptions about the task. For example, many subjects would display the gamblers fallacy: if red on a roulette wheel has come up in every one of the last 4 trials, it must be blacks turn; i.e., they tend to believe that chance is somehow self correcting. In another example, subjects were presented with a probabilistic task - say choosing between two alternatives where reward has been placed behind one of them with unequal probability. Often subjects will match the probability that they observe with their choices, instead of choosing the door with the highest chance of reward (probability matching, [Vulkan, 2000]).

On the contrary, 'low level' perceptual processing, as has mostly been described herein, has been shown during the last 20 years to be near optimal in humans - i.e., subjects' responses are close to following the optimal rule given by Bayes' theorem [Jacobs, 1999, Kording and Wolpert, 2004, Ernst and Banks, 2002, Battaglia et al., 2003]. This optimal use of information may even extend down to the single-neuron level: cells in the human retina lower their membrane potential due to receiving a single photon (or at least very few) [Rieke et al., 1999]; H1 cells in the fly visual system are able to change their response characteristics almost instantaneously based on the distribution of their inputs [Fairhall et al., 2001].

However recent findings have started questioning the divide between perceptual and cognitive processing, as a better understanding of experimental conditions are being developed. E.g. the probability matching problem may not apply when subjects have received plenty of training or are highly motivated [Shanks et al., 2002]. Other studies have shown how subjects are capable of inferring properties of everyday situations based on very limited information, such as the expected future lifetime of a 60 year old man or the expected gross income of a movie given its opening weekend [Griffiths and Tenenbaum, 2006].

The current results fit nicely into this stream, showing that the Bayesian inference that the brain uses for simple perceptual stimuli is structurally similar to the causal inference model that has been shown to be used under more cognitive tasks. If the brain is performing a type of causal inference while engaging in simple perceptual tasks then this naturally breaks down the divide between 'high level' and 'low level' processing, leaving behind a more natural gradual progression.

9.4 Bayes in the brain

Bayesian modeling of a subjects' performance on a behavioral task does in no way prove that the brain actually uses Bayes' rule for its computations. It is quite possible that, as in the Kahneman and Tversky experiments, it is actually utilizing a type of heuristics, albeit one that is close enough to the Bayesian result that we are unable to differentiate the two with our experiment. A much better test would be if we could find direct evidence for Bayesian processing in the nervous system, whether by single neuron recording, fMRI, etc. Although a few studies have shown results compatible with Bayesian processing [Wallace et al., 1998, Gold and Shadlen, 2003, Hampton et al., 2006], no study so far has been able to examine this conclusively.

The problem with testing this is partly a theoretical one: how can we design an experiment that can clearly differentiate whether a set of neurons are performing a 'Bayesian calculation' or not? In order to answer this question we need to have a good theoretical understanding of how a set of neurons would be able to encode a Bayesian computation.

It is useful to contrast Bayes' with how a maximum likelihood (ML) approach would work in the brain. Given, say, a visual stimulus, the ML would find the maximum of the incoming signal and pass this value along to the next step of processing. A Bayesian observer, on the other hand, would take the incoming stimulus, multiply it with the expected distribution (flat or not), and pass along this new distribution to the next processing step. It would therefore also encode its uncertainty about the input and leave any decisions for the later stages (this is similar to the idea of belief propagation [Rao, 2004]). How such encoding would actually happen is far from trivial.

Several studies in the last few years have started to answer these questions [Deneve et al., 2001, Ma et al., 2006, Anastasio et al., 2000]. If we imagine probability distributions encoded as the tuning curve of a set of neurons, each with its own preferred stimulus, we still have the problem of how to encode the multiplication of two such distributions, an operation that is notoriously difficult [Koch, 1999].

One clever approach has recently been presented [Ma et al., 2006] that takes advantage of the Poisson-like firing properties of single neurons. If encoding Gaussian distributions as response functions of a range of single neurons, each with different preferred stimulus, but all firing in a Poisson-like way, the peak of the firing rate is then proportional to the inverse

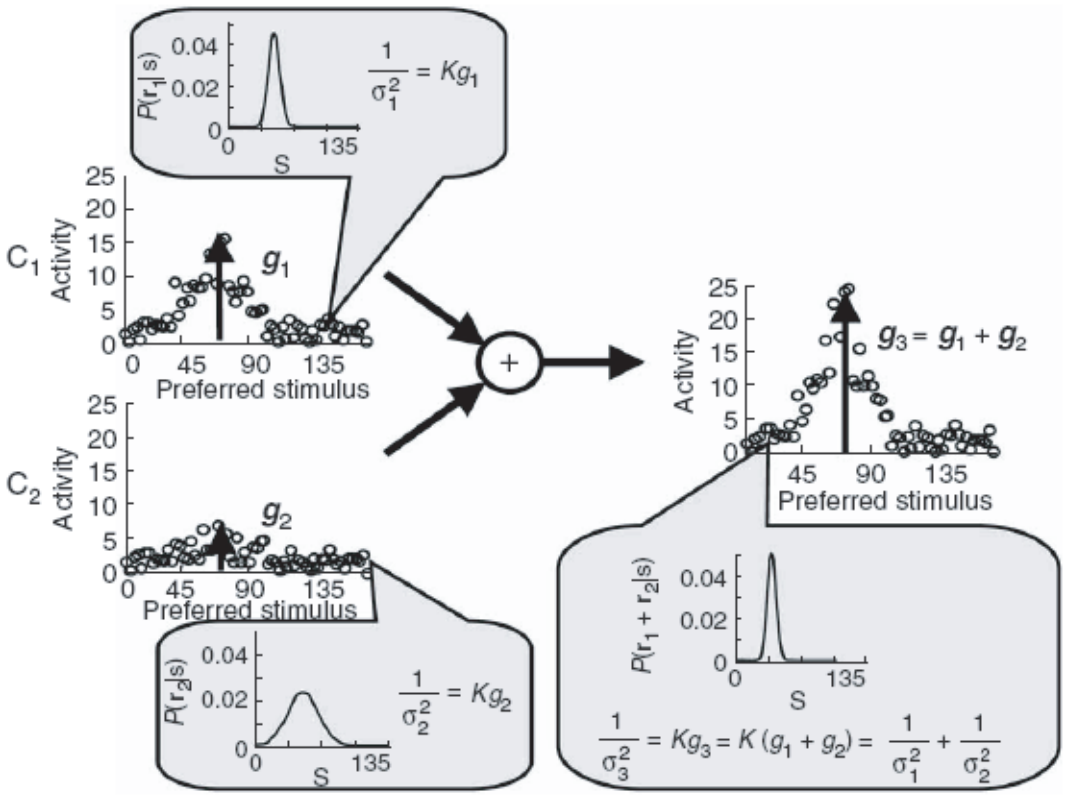


Figure 9.2: (From [Ma et al., 2006].) If a probability distribution is encoded by the population activity of a set of neurons using Poisson firing, then adding the firing rates together is, to a decoder, equivalent to multiplying the two distributions together.

variance of the Gaussian. Adding two firing patterns, say from different stimuli, gives you a new firing distribution with a peak that is now the sum of the two previous peaks, and therefore equal to the sum of the inverse variances. This is exactly the result for variance that we expect when multiplying two Gaussian distributions, $1/\sigma_N^2 = 1/\sigma_1^2 + 1/\sigma_2^2$. We can therefore perform the multiplication of two distributions by simply adding together the firing rates of Poisson-neurons.

This could conceivably be used to simulate more complex situations, as the ones presented in this thesis and further work on this subject should lead to direct predictions testable with, e.g., single electrode recordings.

In the Appendix we will also mention an experiment that is related to this problem of detecting signals relevant for Bayesian encoding in the brain.

9.5 Summary

It is easy get caught by the 'lingua franca' of the literature and start talking about whether or not 'the brain is optimal' [1]. However optimality should more appropriately be seen as an unattainable goal, to be used for comparison. It seems unlikely the nervous system would be utilizing Bayes' theorem for the beauty of the principle itself, instead it will tend to use 'whatever works.' Passing along not just a single value (ML), but the entire computed probability distribution (Bayes), would seem to put a higher demand on the system. If a certain subsystem - say visual detection - has a high enough evolutionary pressure to require a close approximation to optimality, nature will find a way. However for other subsystems, without such a high need for performance, the system will likely use a simpler, less computationally demanding approximation.

So is the human brain optimal?[2] Undoubtedly for some subsystems, but as research continues, using psychophysics, fMRI, single electrode recordings, EEG, and whatever else we can think of, we should find out just how deep the rabbit hole goes.

[1] The author acknowledges occasionally making such statements himself.
[2] Like this.

Appendix A

Future Work: Bayesian Learning

One question we have so far not not dealt with is the problem of learning. The previous studies all assumed that subjects would have to base their estimates on very brief stimuli, with a flat distribution making it very difficult to learn the statistical properties of the stimulus. Here we give an example of how Bayesian learning could be utilized in the human brain.

A.1 Theory

Assume we wish to learn about the property X given some input V. The Bayesian optimal way to derive it is again given by the posterior:

$$P(X|V) = P(V|X)P(X)/P(V) \tag{A.1}$$

where $P(V|X)$ is the likelihood and $P(V)$ is the prior probability of V.

As we saw in equation (2.9) the estimated mean of the posterior can then be written as

$$\hat{x} = \alpha\hat{x}_{prior} + \gamma\hat{V} \tag{A.2}$$

where

$$\alpha = \sigma_V^2/(\sigma_V^2 + \sigma_P^2)$$
$$\gamma = \sigma_P^2/(\sigma_V^2 + \sigma_P^2).$$

σ_V is the width of the likelihood function, σ_P is the width of the prior distribution.

The variance of X becomes:

$$\sigma_X^2 = \left(\frac{1}{\sigma_V^2} + \frac{1}{\sigma_P^2}\right)^{-1}. \tag{A.3}$$

For simplicity of notation, we will refer to $\hat{x}$ as simply x.

For a situation where we have multiple measurements of the same quantity, the posterior probability can be considered the prior for the next measurement [Berger, 1980] so that we instead have an updating rule:

$$x_{t+1} = \alpha x_t + \gamma r_t. \tag{A.4}$$

Since $\alpha = (1 - \gamma)$ we can reduce it to:

$$x_{t+1} = x_t + \gamma(r - x_t). \tag{A.5}$$

We therefore have the following updating rule:

$$x_{t+1} = x_t + \gamma e_t \tag{A.6}$$

where we will refer to the second term as the 'error' term:

$$e_t = r - x_t. \tag{A.7}$$

This is equivalent to a Rescorla Wagner learning rule used in reinforcement learning [Sutton and Barto, 1998], with learning rate

$$\gamma = \sigma_X^2/(\sigma_V^2 + \sigma_P^2). \tag{A.8}$$

After one time step we have

$$\sigma_{X_1}^{-2} = \sigma_V^{-2} + \sigma_P^{-2}, \tag{A.9}$$

after two time-steps we have

$$\sigma_{X_2}^{-2} = \sigma_V^{-2} + (\sigma_V^{-2} + \sigma_P^{-2}), \tag{A.10}$$

and we can easily see that this converges to

$$\sigma_{X_t}^{-2} = t * \sigma_V^{-2} + \sigma_P^{-2}. \tag{A.11}$$

If we set our prior initial variance $\sigma_P^2 = \sigma_V^2$ we get the simple form

$$\gamma = \sigma_V^2/(t+1). \tag{A.12}$$

Let's see how to apply this to a specific problem we have recently started working on.

A.2 Gaze bias

One of the object categories we may be the most familiar with, and the most capable of recognizing is human faces. Whether due to long-term exposure or due to some specific genetic predisposition we are exceptionally adept at classifying and recognizing faces after very short exposure. The social importance of being able to read and analyze faces is undoubtedly part of the reason for this. It is therefore not too surprising that we also form very specific opinions of liking/disliking faces after very short times. In a study by Shimojo et al. [Shimojo et al., 2003] they examined how humans create such preference formation. Subjects were shown two faces on a computer monitor, and were told to choose which face they preferred while their eye movements were recorded.

At least two interesting results came out of this. Firstly subjects were able to perform this task fairly quickly (within 3 - 5 seconds), but had a clear tendency, shortly before making their decision, to gaze at the image they would eventually choose to prefer (see Figure A.1). I.e., they have a bias toward gazing at the eventually preferred image, but only just before making their decision. However if subjects are asked to merely judge the roundness of the face, this bias diminishes significantly.

Possibly more interesting was the second result, that the decision itself could be manipulated. In a second experiment, Shimojo et al. would alternately show two images before

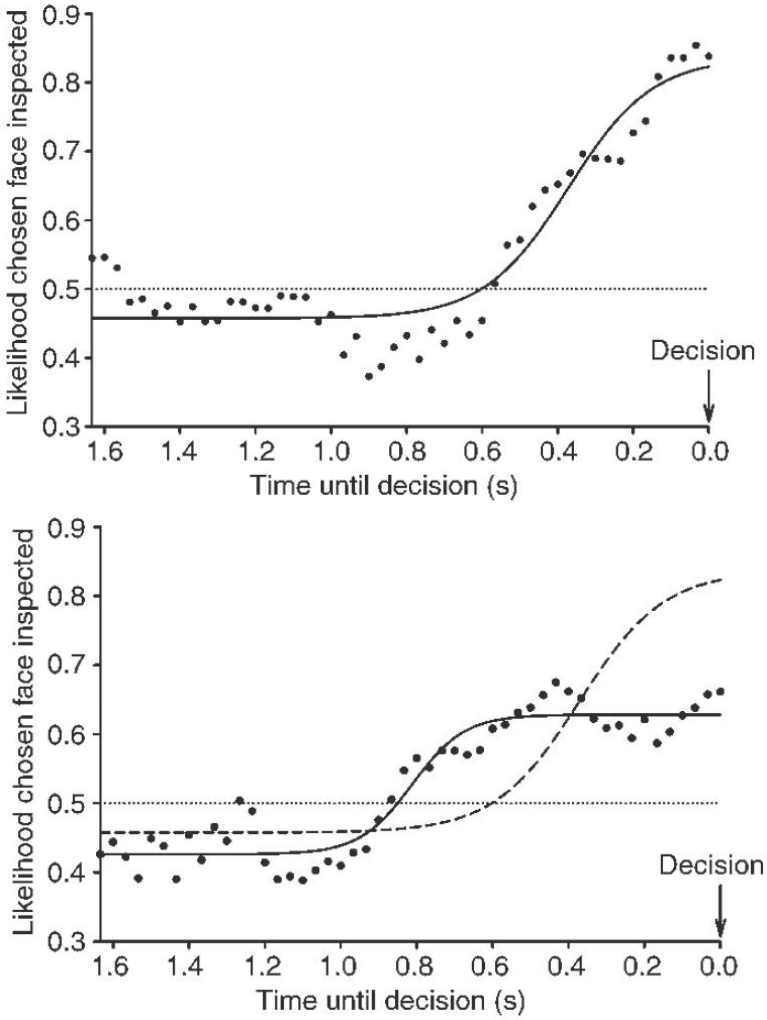


Figure A.1: Example of the gaze bias shown by subjects before making a decision. The upper figure is when subjects were asked to judge which face they liked more; in the lower figure they were asked to judge the roundness of the faces. Curve for the preference task is also plotted in the lower figure (dotted). (From [Shimojo et al., 2003].)

asking subjects to make a decision. By varying the length of exposure to each image they could cause subjects to be more likely to prefer the image presented for longer time. This second bias is therefore a decision bias caused by the exposure to the images.

The question which concerns us here is whether we can use the statistical optimality principle to analyze this problem. The central assumption is that subjects derive utility from watching these images, i.e., that they potentially consider viewing these pictures pleasurable. Several studies have found that animals [Deaner et al., 2005] are willing to accept a decrease in payments (money or juice) in exchange for viewing pictures that they would consider pleasurable, for humans this is well known.

From an optimal gain perspective, the subjects therefore have a trade-off: They want to perform the task as well and fast as possible, which requires exploring the two images - in order to get their monetary reward, but they also wish to derive utility from the viewing of the pictures. Naturally this causes the two images to not be sampled equally, hence we have the basis for a bias. This becomes an instance of an exploration versus exploitation problem, as has been studied intensively in the one armed bandit problem in economical theory [Kelley, 1974, Sutton and Barto, 1998].

For the model we make the following assumptions:

- Subjects derive utility r, from viewing the images.
- However, they base their decisions on the estimated values for the left and right image, x_L and x_R.
- Subjects update their evaluation of the images using the Bayesian approach from above, i.e., using

$$x_{i,t+1} = x_{i,t} + \gamma e_t \tag{A.13}$$

with $e_t = r_i - x_{i,t}$ and $\gamma = 1/(1+t)$.

We model their choices as governed by a soft-max function based on the expected value of the images. The probability of choosing the left image is therefore

$$P(L) = \frac{\exp(\beta * x_L)}{(\exp(\beta * x_L) + \exp(\beta * x_R))}. \tag{A.14}$$

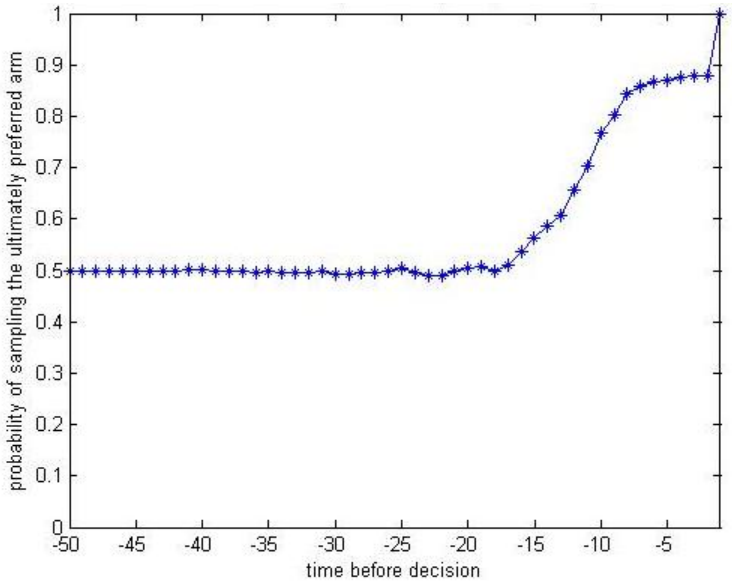


Figure A.2: An example of the average output of the model given a certain parameter set. (Compare with Figure A.1.)

To model the final decision, we assume that a decision is made when either $\sum r_L$ or $\sum r_R$ reaches a threshold θ. One simulation of this model would consist of a series of left and right decisions, but by running this model a large number of times, we can create an average time series of 'gaze-locations' corresponding to the 'gaze-locations' in the experimental task. Since the model assumes a simple cumulative threshold for the final decision, it is directly compatible with subjects choosing the image they have been exposed to longer, as the cumulative value will tend to be higher. Further, as shown in Figure A.2, the time series is qualitatively similar to the gaze bias results from the study of Shimojo et al.

A.3 Work to be done

Clearly we are using an overly simplified model to understand this task, but as a starting point it has certain advantages: it is able to explain the second gaze-bias (bias toward choosing the more seen), since it is built in to the model and it shows the same behavior as the data with regard to the first bias (bias toward looking at the eventually chosen image). Questions to consider include how to test such a model, how to model correlations in the experimental data, and how to model presentation of non-face images.

Appendix B

Future Work: fMRI

Everything so far has dealt with questions regarding behavior, to what degree is human behavior optimal? Another way of approaching the general subject of Bayesian processing in the nervous system is to ask whether we can find evidence of Bayesian computation by recording from the brain. As described in the discussion, no studies have been able to conclusively detect signs of Bayesian processing. We have begun a study that using fMRI, indirectly asks if the cortex does perform optimal processing.

B.1 Optimal gambling

As discussed earlier (see e.g., Chapter 9), cognitive problem solving has often been found to be sub-optimal, whereas perceptual tasks can be performed close to optimally.

Recently Daw et al. [Daw et al., 2005] have proposed a theory of competition between two subsystems of the cortex; a 'habitual' system consisting of parts of the striatum, and a 'model-based' system located in the pre-frontal cortex. Several previous studies have implicated these areas in separate learning and reward schemes [Packard and Knowlton, 2002, O'Doherty et al., 2004, Schultz et al., 1997].

A habitual system would be model free and would only need to do simple updating of a few variables, Pavlovian conditioning would be an example. In contrast, a model-based system would take all the available information into account and update all its variables accordingly. Following the idea by Daw et al., these systems separately and simultaneously update the subjective values of the options for the subject. Using an approach derived from Bayesian inference, the action represented by the system with the smaller uncertainty gets chosen, and can therefore vary between tasks.

The model-based approach would be more efficient, but might require more resources, whereas the habitual system is easy and cost effective to implement, but more prone to mistakes. Confirming the presence of these separate systems within a single task and being able to observe the switching between them would provide strong support for this idea.

B.2 Gambling task

For these reasons we set up an experiment that allowed us to test the contrast between subjects habitual and cognitive (model-based) behavior. Subjects were shown three doors, behind which an amount of money had been hidden. Subjects were instructed to order the doors in order of the likelihood of their payout, but were not told how to estimate this likelihood. They were however informed that the likelihood of a door hiding the money was constant across rounds and that the sequence of the doors was chosen before the experiment (i.e., independent of their decisions). The instructions were purposefully vague in order for us to be able to observe different types of behavior.

If the money was behind the door subjects had as their first choice, they would receive 50 cents, if behind their second choice 0 cents and if behind their third choice they would lose 50 cents. It was therefore in their interest to pay attention to both their first and second choice. The likelihood of the money being assigned to each door was fixed during each block of trials; subjects played 3 blocks of each 40 trials.

Before testing the subjects inside the fMRI scanner we asked them to fill out a questionnaire, to make sure they understood the concepts, and allowed them to play the game very briefly, to make sure they understood the mechanics. After the scan we asked the subjects to fill out another questionnaire asking them about their strategy and perceptions of the game.

B.2.1 Modeling

For this task there is a very simple optimal solution, given by the realization that there is an underlying probability distribution which needs to be estimated. The optimal solution can be derived from Bayesian inference and the likelihood associated with door i is

$$P_i = \frac{n_i + 1}{\sum_i n_i + 1} \tag{B.1}$$

where n_i is the observed number of times that door i has had the money. However this is also very close to an intuitively simple 'counting rule,' with a prior distribution. Given the estimated likelihood, the best strategy is to pick the door with the highest likelihood as your first choice, and the second highest as your second choice. The expected reward can then be calculated as

$$E < R >= \max(p_i) * 50c - \min(p_i) * 50c = (\max(p_i) - \min(p_i)) * 50c. \tag{B.2}$$

We will assume that this is indeed what the subjects wish to maximize.

However there is a suboptimal way of approaching this problem, given by reinforcement learning (RL). RL has been found to be an excellent way to model value updating in a large number of studies [Sutton and Barto, 1998, O'Doherty et al., 2003, McClure et al., 2003]. The model is here similar to the one in Appendix A.

Here we assume that subjects choices at time t constitute a set of responses $C_{i,t}$, the value of which gets updated using reinforcement learning. There are therefore 6 sets of response possibilities for their first and second choices: $C_{1,t} = (1,2), C_{2,t} = (1,3), C_{3,t} = (2,1), C_{4,t} = (2,3), C_{5,t} = (3,1), C_{6,t} = (3,2)$. The value of each of these gets updated whenever they are used, as follows

$$x_{i,t+1} = x_{i,t} + \gamma e_t \tag{B.3}$$

with $e_t = r_i - x_{i,t}$ and $\gamma = 1/(1+t)$.

This may seem like a very complex model, but from the subjects' viewpoint it does not include any assumptions about underlying structure. Given a response, subjects receive a reward and update their values based on that. In contrast, the optimal probabilistic model above (equation (B.1)) requires a specific understanding of how the money is distributed (fixed distribution, etc.).

An observer utilizing all the available information would therefore perform according to equation (B.1), whereas someone making no assumptions might rely on the reinforcement learning model.

Given subjects' behavioral choices, we wish to compare the trial-based expected reward, using both the optimal and sub-optimal model, and correlate these with the activity in areas like the ventral striatum, prefrontal cortex, etc., using fMRI.

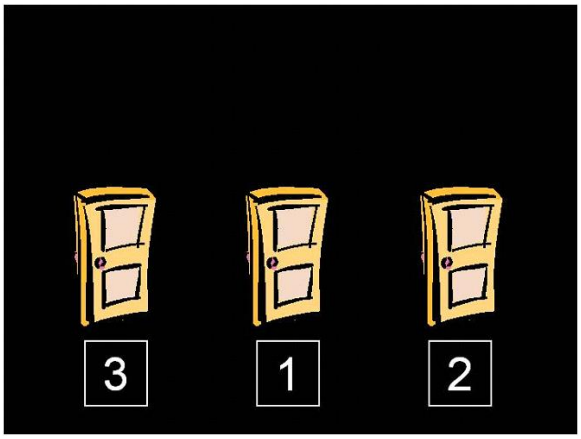


Figure B.1: The decision screen in our gambling task, after a subject has made his/her choices - here preferring the center door and picking the right door as the second choice.

B.2.2 Behavioral results

27 subjects were scanned while performing the three blocks of the game. One subject did not complete the third block and the data was therefore discarded.

A large range of behaviors were observed between subjects. Some subjects (n=9) figured out the optimal strategy either consciously (n=6 as self reported) or unconsciously (n=3), and made close to the optimal amount of money (17 dollars). This is the amount of money someone playing the optimal strategy would be making, however notice that nothing prevents someone playing sub-optimally to make more money in the short run, due to sheer luck. In contrast, several subjects (n=8) would play in an inconsistent way, e.g., preferring the left door in one round and the right door in the next round, even at late rounds in a block. A minority of these subjects (n=1) would afterwards report trying to outsmart the computer or looking for patterns (n=4), and had clearly misunderstood the instructions.

B.3 Work to be done

The results discussed here only relate to the behavioral results, we still need to analyze the MRI data which requires a good understanding of how to model the HRF (hemodynamic response function), given the task. Preliminary results are, however, encouraging - e.g., a simple contrast between rounds where subjects won versus rounds where they lost, shows clear activation in typical reward related areas (see Figure B.2).

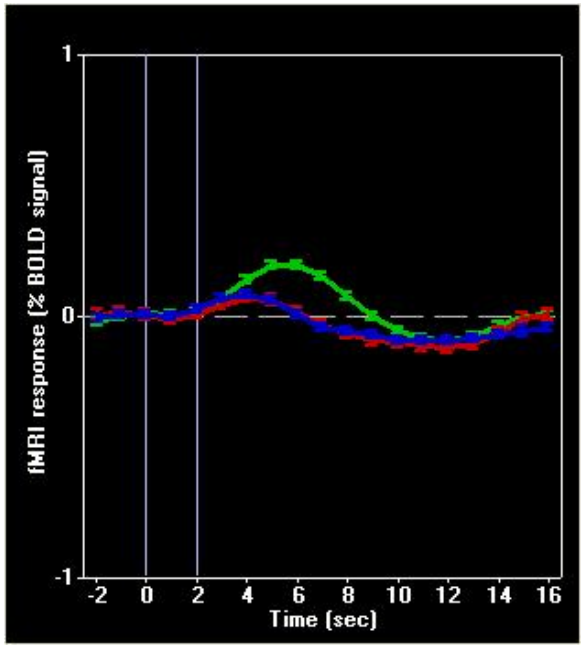


Figure B.2: Average activation in part of ventral striatum after receiving reward (green), no reward (blue), and losing money (red). The result is revealed at time zero.

Furthermore, this is just a small step towards testing the optimal processing hypothesis in the human brain, and an indirect step at that. As discussed previously, in order to really test whether the brain utilizes Bayesian processing we need a better theoretical understanding of how the computation would be presented in the brain and what testable predictions this would create.

Appendix C

Mathematics and Methods

This chapter goes through some of the results, mathematics, and methods for which there was no room in the main text.

C.1 Causal inference

C.1.1 Details of the generative model

The Bayesian generative-model-based approach necessitates fully specifying how the variables one is interested in are interrelated in a statistical fashion. From such a specification there is then a unique solution of optimal inference. The specification of all assumptions and the fact that the inference procedure is devoid of any parameters makes it easy to compare different approaches within the Bayesian framework.

The generative model is specified as follows:

Determine if there is one cause ($common$) versus two causes ($\neg common$) by drawing from a binomial distribution with $P(common) = p_{common}$.

- if one cause ($common$) draw position X from a normal distribution $N(0, \sigma_{pos})$, where $N(\mu, \sigma)$ stands for a normal distribution with mean μ and standard deviation σ. Then draw V from $N(x, \sigma_V)$ and A from $N(x, \sigma_A)$.

- if two causes ($\neg common$) draw position X_V and X_A each independently from $N(0, \sigma_{pos})$.

Then draw $V$from $N(X_V, \sigma_V)$ and A from $N(X_A, \sigma_A)$

C.1.2 Estimating the probability of a common cause

An ideal observer is then faced with the problem of doing inference about the causal structure, whether there is one cause (X) or two causes (X_V and X_A). This inference is done

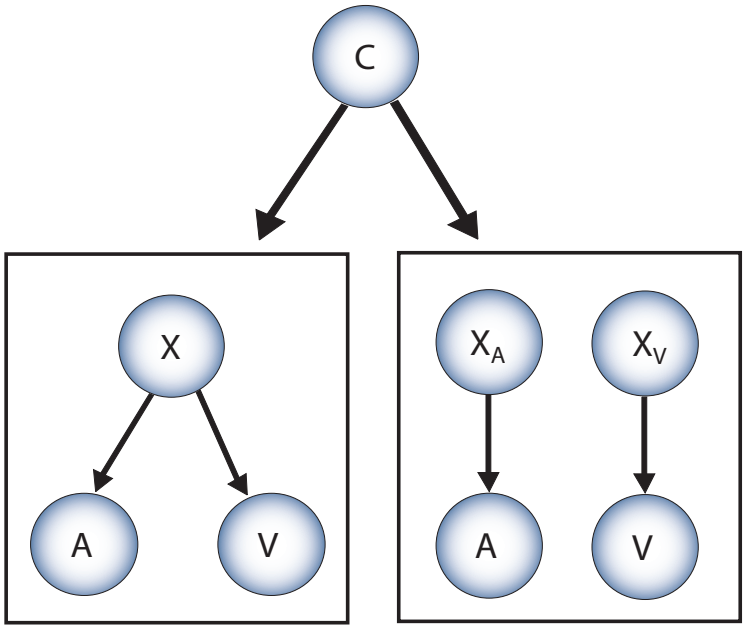


Figure C.1: An alternative graphical model representation of the causal inference model. For every trial the positions of X_V, X_A, X, and the switching variable common are drawn randomly according to their distributions. If common is true then X_V and X_A are ignored, otherwise X is ignored. The positions V and A are always corrupted by additive Gaussian noise with width σ_V and σ_A, respectively.

optimally using Bayes' rule:

$$P(common|V,A) = \frac{P(V,A|common)P(common)}{P(V,A)}. \tag{C.1}$$

Here $P(V,A)$ must be chosen so that $P(common|V,A)$ and $P(\neg common|V,A)$ add to 1, as we are dealing with probabilities. We thus obtain:

$$P(common|V,A) = p_{common}P(V,A|common) + (1-p_{common})P(V,A|\neg common). \tag{C.2}$$

For $P(V,A|common)$ we obtain

$$P(V,A|common) = \int P(V|X)P(A|X)P(X)dX. \tag{C.3}$$

All three factors in this integral are Gaussian distributions allowing an analytic solution. For $P(V,A|\neg common)$ we realize that V and A are independent of each other and thus obtain:

$$P(V,A|\neg common) = \int P(V|X_V)P(X_V)dX_V \int P(A|X_A)P(X_A)dX_A. \tag{C.4}$$

Again, as all these distributions are assumed to be Gaussian, we obtain an analytic solution. We assume that when $P(common|V,A) > 0.5$ the model reports a common cause for comparison with experimental data in Section 5.4. The estimation of a common cause, which we have solved in this section, amounts to a Bayesian model selection problem.

C.1.3 Optimally estimating the position

When estimating the position of the visual target we assume that making an error about the position of x in the case of a common cause is just as bad as making an error in the estimate of x_V, and likewise for the position of the auditory target. We assume that the cost in this estimation process is the mean expected deviation:

$$Cost = E[P(common|V,A)(x_{est}-x)^2 + P(\neg common|V,A)(x_{est}-x_{AorV})^2] \tag{C.5}$$

where $E[]$ is the expected value, x_{est} is the possible estimate and x_{VorA} is x_V in the case of a visual task and x_A in the case of an auditory task. We observe that when the cost

function is quadratic, the optimal estimation problem reduces to the problem of finding the mean of the posterior. The solution that minimizes the expected error is:

$$\hat{x} = P(common|V, A)\hat{x}_{common} + (1 - P(common|V, A))\hat{x}_{\neg common} \tag{C.6}$$

where $\hat{x}$ is the best estimate, and $\hat{x}_{common}$ and $\hat{x}_{\neg common}$ are the best estimates we would have if we were certain about a common cause or independent causes, respectively. The solutions of $\hat{x}_{common}$ and $\hat{x}_{\neg common}$ are obtained by linearly weighing the different cues proportional to their inverse variances. This solution follows from the fact that the posterior is a product of Gaussian distributions and thus a Gaussian itself. In this case it is possible to analytically solve for the maximum a posteriori solution which coincides with the mean squared error solution. The optimal solution when having one cue and a prior is:

$$\hat{x} = \alpha\hat{x}_{sensory} + (1 - \alpha)\hat{x}_{prior} \tag{C.7}$$

where $\hat{x}$ is the best estimate, $\hat{x}_{sensory}$ the position perceived by the senses (from the likelihood), and $\hat{x}_{prior}$ the mean of the prior. For the weight α we obtain $\alpha = \sigma^2_{prior}/(\sigma^2_{prior} + \sigma^2_{sensory})$, where σ_{prior} is the width of the prior and $\sigma_{sensory}$ is the uncertainty in the sensory information. The optimal solution when having two cues is:

$$\hat{x} = \frac{(\sigma^2_A\sigma^2_{prior}x_V + \sigma^2_V\sigma^2_{prior}x_A + \sigma^2_V\sigma^2_A x_{prior})}{\sigma^2_V\sigma^2_A + \sigma^2_A\sigma^2_{prior} + \sigma^2_V\sigma^2_{prior}}. \tag{C.8}$$

While these equations are linear, $P(common|V, A)$ is nonlinear. For that reason the optimal strategy is no longer a linear strategy. The traditional way of studying the system, by only measuring gains and linear weights, thus is not sufficient to fully describe a system that does causal inference. To link the model with our data, we assume that people press the button which is associated with the position closest to $\hat{x}$. This amounts to only allowing estimated values x_{est} that are out of the set: $-10, -5, 0, 5, 10$ degrees. To link the model with the data on causal inference of Figure 5.4 in Chapter 5 we assume additional motor noise with width σ_{motor} that is perturbing the estimated position x_{est}.

C.1.4 Derivation of causal inference model

$$< X_A >=< P(X_A|A, V) > \tag{C.9}$$

$$< X_A >=< \sum P(X_A, common|A, V) > \tag{C.10}$$

$$< X_A >=< \sum P(X_A|A, V, common)P(common|A, V) > \tag{C.11}$$

$$< X_A >= \sum P(common|A, V) < P(X_A|A, V, common) > \tag{C.12}$$

$$< X_A >= P(common|A, V) < P(X_A|A, V, common) > +P(\neg common|A, V) < P(X_A|A, V, \neg common) \tag{C.13}$$

$$\hat{x}_A = P(common|A, V)\hat{x}_{A,common} + P(\neg common|A, V)\hat{x}_{A,\neg common} \tag{C.14}$$

C.1.5 Formulas for causal inference

In order to calculate $\hat{X}$ we need to estimate $P(common|A, V)$:

$$P(common|A, V) = \frac{P(A, V|common)P(common)}{P(A, V)}. \tag{C.15}$$

P(common) is one of the parameters we are fitting, $P(A, V|common)$ is given by

$$P(A, V|common) = \int_x P(A, V|X)P(X)dx = \int_x P(A|X)P(V|X)P(X)dx \tag{C.16}$$

$$\begin{aligned} P(A, V|\neg common) &= \int_{X_V, X_A} P(A, V|X_A, X_V)P(X_A, X_V)dX_V dX_V \\ &= \int_{X_A} P(A|X_A)P(X_A)dX_A \int_{X_V} P(V|X_V)P(X_V)dX_V. \end{aligned}$$

If we assume that both likelihoods and priors are Gaussian distributed, we can get

analytical expressions for these integrals. For the integration of two Gaussian distributions:

$$\int_{X_A} P(A|X_A)P(X_A)dX_A = \int_{X_A} \frac{1}{\sqrt{2\pi\sigma_A^2}} \exp\left(-\frac{1}{2}\frac{(X_A-\mu_A)^2}{\sigma_A^2}\right) \frac{1}{\sqrt{2\pi\sigma_X^2}} \exp\left(-\frac{1}{2}\frac{(X_A-\mu_X)^2}{\sigma_X^2}\right) dX_A = \int_{X_A} \frac{1}{2\pi\sigma_A\sigma_X} \exp\left(-\frac{1}{2}(\frac{(X_A-\mu_A)^2}{\sigma_A^2} + \frac{(X_A-\mu_X)^2}{\sigma_X^2}\right) dX_A.$$

Eventually we get:

$$\int_{X_V} P(A|X_A)P(X_A)dX_V = \frac{1}{\sqrt{2\pi(\sigma_A^2+\sigma_X^2)}} \exp\left(-\frac{1}{2}(\frac{(\mu_X^2-\mu_A)^2}{\sigma_A^2+\sigma_X^2}\right). \tag{C.17}$$

A solution for the integral of three Gaussian distributions can be found in similar ways.

C.2 'Super' model

C.2.1 Derivation

$$\hat{x}_V = P(cor|A,V)\hat{x}_{V,cor} + P(cor|A,V)\hat{x}_{V,cor} \tag{C.18}$$

We can use the general rule for the product of two normal distributions; the new mean is:

$$x_{\hat{V,cor}} = \left(\Sigma_{AV}^{-1} + \Sigma_{cor}^{-1})\right)^{-1} (\Sigma_V^{-1}\hat{V} + \Sigma_{prior}^{-1}\mu_{prior}) \tag{C.19}$$

where we have used vector notation so that Σ is a covariance matrix and μ is the mean vector. Similarly:

$$\hat{x}_{V,\neg cor} = \frac{\hat{x}_V/\sigma_V^2 + \mu_X/\sigma_X^2}{(1/\sigma_X^2 + 1/\sigma_V^2)}. \tag{C.20}$$

C.2.2 Formulas for super model

We now assume that the prior is a sum of 2 normal distributions, one with full covariance and one with 0 covariance. The second part is similar to the second part for the causal inference prior.

In order to calculate $\hat{X}$ we need to estimate $P(cor|A,V)$:

$$P(cor|A,V) = \frac{P(A,V|cor)p_{cor}}{P(A,V)}. \tag{C.21}$$

p_{cor} is one of the parameters we are fitting, $P(A,V|cor)$ is given by:

$$P(A,V|cor) = \int_{X_V}\int_{X_A} P(A,V|X_A,X_V)P(X_A,X_V|cor)dX_V dX_A$$
$$P(A,V|\neg cor) = \int_{X_V}\int_{X_A} P(A|X_A)P(V|X_V)P(X_A,X_V|\neg cor)dX_V dX_A.$$

$P(X_A,X_V|cor)$ is a 2-dimensional Gaussian distribution with full covariance, so we need to find the integral of the product of 2 2D Gaussian distributions.

$$P(A,V|cor) = \frac{1}{2\pi\sqrt{|\Sigma_{AV}| + |\Sigma_{cor}|}} \exp\left(-\frac{1}{2}(\mu_{cor} - \mu_{AV})(\Sigma_{cor} + \Sigma_{AV})^{-1}(\mu_{cor} - \mu_{AV})\right). \tag{C.22}$$

For $P(A,V|\neg cor)$ we can use the previous result:

$$\begin{aligned} P(A,V|\neg cor) &= \int_{X_V,X_A} P(A,V|X_A,X_V)P(X_A,X_V)dX_V dX_A \\ &= \int_{X_A} P(A|X_A)P(X_A)dX_A \int_{X_V} P(V|X_V)P(X_V)dX_V. \end{aligned}$$

C.3 Relation between causal and 2D Bayesian model

We wish to examine the relationship between the Bayesian model with a 2-dimensional prior and the causal inference model.

We start with our original posterior

$$P(X_A, X_V|A, V) = \frac{P(A|X_A)P(V|X_V)P(X_A, X_V)}{P(A, V)}. \tag{C.23}$$

We will assume a specific shape for the prior,

$$P(X_A, X_V) = \sum_c P(X_A, X_V|C)P(C) = P(C)P(X|C) + P(\neg C)P(X_A)P(X_V) \tag{C.24}$$

where C is a binary variable (short for common or uncommon) $P(X|C)$ is a perfectly diagonal distribution with $P(X|C) = P(X_A, X_V|common) = \delta(X_A - X_V)P(X_V = X_A)$, and $P(X|\neg C) = P(X_V)P(X_A)$.

$$P(X_A, X_V|A, V) = P(A|X_A)P(V|X_V)\sum_c P(X_A, X_V|C)P(C)/P(A, V) \tag{C.25}$$

Assume for now that we only care about X_A

$P(X_A|A, V) = \int P(X_A, X_V|A, V)dX_V$

$$= \int P(A|X_A)P(V|X_V)\sum_c P(X_A, X_V|C)P(C)/P(A, V)dX_V \tag{C.26}$$

$$P(X_A|A, V) = \int P(X_A, X_V|A, V)dX_V = \int P(A|X_A)P(V|X_V)P(X_A, X_V)/P(A, V)dX_V. \tag{C.27}$$

Take the mean

$$< X_A >= \int X_A P(X_A|A, V)dX_A = \int\int X_A P(A|X_A)P(V|X_V)P(X_A, X_V)/P(A, V)dX_V dX_A. \tag{C.28}$$

Split the prior

$$P(X_A, X_V) = \sum_C P(X_A, X_V|C)P(C) \tag{C.29}$$

$$< X_A >= \int\int X_A P(A|X_A) P(V|X_V) \sum_C P(X_A, X_V|C) P(C)/P(A,V) dX_V dX_A \quad \text{(C.30)}$$

using that

$$P(C)/P(A,V) = P(C|A,V)/P(A,V|C) \quad \text{(C.31)}$$

$$< X_A >= \sum_C P(C|A,V)/P(A,V|C) \int\int X_A P(A|X_A) P(V|X_V) P(X_A, X_V|C) dX_V dX_A \quad \text{(C.32)}$$

$$\begin{aligned} < X_A >= & P(C=1|A,V)/P(A,V|C=1) \int\int X_A P(A|X_A) P(V|X_V) P(X_A, X_V|C=1) dX_V dX_A + \\ & P(C=2|A,V)/P(A,V|C=2) \int\int X_A P(A|X_A) P(V|X_V) P(X_A, X_V|C=2) dX_V dX_A \end{aligned}$$

using that

$$P(X_A, X_V|C=1) = \delta_{X_A - X_V} P(X_A|C=1) \quad \text{(C.33)}$$

and using that

$$P(X_A, X_V|C=2) = P(X_A|C=2) P(X_V|C=2) \quad \text{(C.34)}$$

$$\begin{aligned} < X_A >= & \frac{P(C=1|A,V)}{P(A,V|C=1)} \int X_A P(A|X_A) P(V|X_A) P(X_A|C=1) dX_A + \\ & \frac{P(C=2|A,V)}{P(A,V|C=2)} \int X_A P(A|X_A) P(X_A|C=2) dX_A \int P(V|X_V) P(X_V|C=2) dX_V \end{aligned}$$

using

$$P(A,V|C=2) = P(A|C=2) P(V|C=2) \quad \text{(C.35)}$$

$$< X_A >= P(C|A,V) \int X_A \frac{P(A|X_A)P(V|X_A)P(X_A|C=1)}{P(A,V|C)} dX_A + P(C|A,V) \int X_A \frac{P(A|X_A)P(X_A|C=2)}{P(A|C=2)} dX_A \int \frac{P(V|X_V)P(X_V|C=2)}{P(V|X=2)} dX_V$$

$$< X_A >= P(C=1|A,V) \int X_A P(X_A|A,V,C=1) dX_A + P(C=2|A,V) \int X_A P(X_A|A) dX_A \int P(X_V|V|C=2) dX_V$$

$$< X_A >= P(C=1|A,V) < X_c > + P(C=2|A,V) < X_c > . \tag{C.36}$$

The calculation is similar for X_V. We can therefore see that the causal inference model is equivalent to our original model with a 2D prior that takes a very specific form, a weighted average of a diagonal matrix, and a decomposable matrix given by the independent priors $P(X_A)$ and $P(X_V)$.

C.4 Head related transfer function

The HRTF is a transfer function that represents the spatial transformation of a stereo audio signal. When convolving an arbitrary stereo signal Y with the specific transfer function H, $f = H * Y$ where $*$ signifies convolution, the signal is transformed in a way similar to the transformation done by the environment on the sound as it moves from a speaker to the subjects' ears. It is therefore used as a substitute for actual speakers, as it can transform the sound as if it was projected from that direction. The function is therefore dependent on the direction environment and the shape of the subjects ear, head, and shoulders. For this reason we chose to measure it directly from subjects individually, as opposed to using a generic one.

The method used was similar to the one employed in MIT's Kemar study[1]. Subjects were outfitted with a pair of in-ear microphones and their heads were placed in a chin rest. Maximum Length Sequences of 15 bits were generated and presented through a single speaker placed in one of the 5 directions we were testing, at the same distance from the subject as the monitor for the later testing. Maximum Length Sequences are ideal for this purpose since the cross correlation between the recorded sequence and the original sequence directly gives the transfer function [Rife and Vanderkooy, 1989]. The sound from the speaker was recorded and the cross correlation with the original signal was calculated. This was repeated 20 times, the resulting transfer functions were aligned and averaged to create a mean transfer function for that location. A new function was calculated for each location and each subject.

[1] http://sound.media.mit.edu/KEMAR.html

Figure C.2: Microphones like this were inserted into the subjects' ears.

References

L. Abbott and P. Dayan. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. Cambridge, MA: MIT Press, 2005.

D. Alais and D. Burr. The ventriloquist effect results from near-optimal bimodal integration. *Curr Biol*, 14(3):257–62, 2004. URL `http://dx.doi.org/10.1016/j.cub.2004.01.029`.

T. J. Anastasio, P. E. Patton, and K. Belkacem-Boussaid. Using bayes' rule to model multisensory enhancement in the superior colliculus. *Neural Computation*, 12(5):1165–1187, 2000. doi: 10.1162/089976600300015547. URL `http://www.mitpressjournals.org/doi/abs/10.1162/089976600300015547`.

R. A. Andersen. Multimodal integration for the representation of space in the posterior parietal cortex. *Philos Trans R Soc Lond B Biol Sci*, 352(1360):1421–8, 1997.

P. W. Battaglia, R. A. Jacobs, and R. N. Aslin. Bayesian integration of visual and auditory signals for spatial localization. *J Opt Soc Am A Opt Image Sci Vis*, 20(7):1391–7, 2003.

T. Bayes. Essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 1764.

J. Berger. *Statistical Decision Theory and Bayesian Analysis*. New York: Springer-Verlag, 1980.

C. Bishop. *Neural Networks for Pattern Recognition*. Oxford, U.K.: Oxford University Press, 1995.

R. Bowen, J. Mallow, and P. Harder. Some properties of the double-flash illusion. *J. Opt. Soc. Am. A*, 4:746–755, 1987.

D. Brainard. The psychophysics toolbox. *Spatial Vision*, 10:433–436, 1997.

M. J. Buehner, P. W. Cheng, and D. Clifford. From covariation to causation: a test of the assumption of causal power. *J Exp Psychol Learn Mem Cogn*, 29(6):1119–40, 2003. URL http://dx.doi.org/10.1037/0278-7393.29.6.1119.

H. H. Bulthoff and H. A. Mallot. Integration of depth modules: stereo and shading. *J Opt Soc Am A*, 5(10):1749–58, 1988.

J. J. Clark and A. L. Yuille. *Data fusion for sensory information processing systems*. Boston: Kluwer Academic Pulishers, 1990.

N. D. Daw, Y. Niv, and P. Dayan. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat Neurosci*, 8(12):1704–11, 2005. URL http://dx.doi.org/10.1038/nn1560.

R. O. Deaner, A. V. Khera, and M. L. Platt. Monkeys pay per view: Adaptive valuation of social images by rhesus macaques. *Curr Biol*, 15:543–48, 2005. URL http://dx.doi.org/10.1016/j.cub.2005.01.044.

S. Deneve, P. E. Latham, and A. Pouget. Efficient computation and cue integration with noisy population codes. *Nat Neurosci*, 4(8):826–31, 2001. URL http://dx.doi.org/10.1038/90541.

M. P. Eckstein, C. K. Abbey, B. T. Pham, and S. S. Shimozaki. Perceptual learning through optimization of attentional weighting: human versus optimal Bayesian learner. *J Vis*, 4 (12):1006–19, 2004. URL http://dx.doi.org/10:1167/4.12.3.

M. Ernst. A bayesian view on multimodal cue integration. In G. M. Knoblich, I. Grosjean, and M. Thornton, editors, *Perception of the human body from the inside out,*, pages 105–131. Oxford University Press, New York, 2006.

M. O. Ernst and M. S. Banks. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–33, 2002. URL http://dx.doi.org/10.1038/415429a.

A. L. Fairhall, G. D. Lewen, W. Bialek, and R. R. de Ruyter Van Steveninck. Efficiency and ambiguity in an adaptive neural code. *Nature*, 412(6849):787–92, 2001. URL http://dx.doi.org/10.1038/35090500.

A. Falchier, S. Clavagnier, P. Barone, and H. Kennedy. Anatomical evidence of multimodal integration in primate striate cortex. *J Neurosci*, 22(13):5749–59, 2002. URL `http://dx.doi.org/20026562`.

E. Fantino. Behavior analysis and decision making. *Journal of the experimental Analysis of behavior*, 69(3):355–364, 1998.

S. Gepshtein, J. Burge, M. O. Ernst, and M. S. Banks. The combination of vision and touch depends on spatial proximity. *J Vis*, 5:1013–1023, 2005.

Z. Ghahramani, D. M. Wolpert, and M. I. Jordan. Computational models of sensorimotor integration. In P. G. Morasso and V. Sanguineti, editors, *Selforganization, computational maps, and motor control*. San Mateo, California: Morgan Kaufmann, 1988.

Z. Gharahmani. *PhD thesis*. Massachussetts Institute of Technology, 1995.

J. I. Gold and M. N. Shadlen. The influence of behavioral context on the representation of a perceptual decision in developing oculomotor commands. *J Neurosci*, 23(2):632–51, 2003.

A. Gopnik, D. M. Sobel, L. Schulz, and C. Glymour. Causal learning mechanisms in very young children: Two, three, and four-year olds infer causal relations from patterns of variation and covariation. *Developmental Psychology*, 37:620–629, 2001.

A. Gopnik, C. Glymour, D. M. Sobel, L. E. Schulz, T. Kushnir, and D. Danks. A theory of causal learning in children: causal maps and Bayes nets. *Psychol Rev*, 111(1):3–32, 2004. URL `http://dx.doi.org/10.1037/0033-295X.111.1.3`.

T. L. Griffiths and J. B. Tenenbaum. Structure and strength in causal induction. *Cognit Psychol*, 51(4):334–84, 2005. URL `http://dx.doi.org/10.1016/j.cogpsych.2005.05.004`.

T. L. Griffiths and J. B. Tenenbaum. Optimal predictions in everyday cognition. *Psychol Sci*, 17(9):767–73, 2006. URL `http://dx.doi.org/10.1111/j.1467-9280.2006.01780.x`.

T. L. Griffiths and J. B. Tenenbaum. From mere coincidences to meaningful discoveries. *Cognition*, 103(2):180–226, 2007. URL `http://dx.doi.org/10.1016/j.cognition.2006.03.004`.

T. L. Griffiths, E. R. Baraff, and J. B. Tenenbaum. Using physical theories to infer hidden causal structure. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society*, 2004.

W. D. Hairston, M. T. Wallace, J. W. Vaughan, B. E. Stein, J. L. Norris, and J. A. Schirillo. Visual localization ability influences cross-modal bias. *J Cogn Neurosci*, 15(1):20–9, 2003. URL `http://dx.doi.org/10.1162/089892903321107792`.

S. T. Hammett, R. A. Champion, P. G. Thompson, and A. B. Morland. Perceptual distortions of speed at low luminance: evidence inconsistent with a Bayesian account of speed encoding. *Vision Res*, 47(4):564–8, 2007. URL `http://dx.doi.org/10.1016/j.visres.2006.08.013`.

A. N. Hampton, P. Bossaerts, and J. P. O'Doherty. The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *J Neurosci*, 26(32):8360–7, 2006. URL `http://dx.doi.org/10.1523/JNEUROSCI.1010-06.2006`.

J. M. Hillis, M. O. Ernst, M. S. Banks, and M. S. Landy. Combining sensory information: mandatory fusion within, but not between, senses. *Science*, 298(5598):1627–30, 2002. URL `http://dx.doi.org/10.1126/science.1075396`.

P. W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81:945–960, 1986.

I. Howard and W. Templeton. *Human Spatial Orientation*. London, UK: Wiley, 1966.

D. Hume. *An Enquiry concerning Human Understanding*. Oxford, UK: Clarendon Press, 1777.

R. A. Jacobs. Optimal integration of texture and motion cues to depth. *Vision Res*, 39 (21):3621–9, 1999.

R. A. Jacobs and I. Fine. Experience-dependent integration of texture and motion cues to depth. *Vision Res*, 39(24):4062–75, 1999.

M. I. Jordan. Graphical models. *Stat. Sci.*, 19:140–155, 2004.

E. Kandel, J. Schwartz, and T. Jessel, editors. *Principles of Neural Science*. McGraz-Hill, 2000.

T. A. Kelley. A note on the bernoulli two-armed bandit problem. *The Annals of Statistics*, 2:1056–1062, 1974.

D. C. Knill. Mixture models and the probabilistic structure of depth cues. *Vision Res*, 43 (7):831–54, 2003.

D. C. Knill and A. Pouget. The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci*, 27(12):712–9, 2004. URL `http://dx.doi.org/10.1016/j.tins.2004.10.007`.

C. Koch. *Biophysics of Computation: Information Processing in Single Neurons (Computational Neuroscience Series)*. Oxford University Press, Inc., New York, NY, USA, 1999.

K. P. Kording and D. M. Wolpert. Bayesian integration in sensorimotor learning. *Nature*, 427(6971):244–7, 2004. URL `http://dx.doi.org/10.1038/nature02169`.

T. Kuhn. *The Structure of Scientific Revolutions*. University Of Chicago Press; (3rd edition 1996), 1962.

M. S. Landy, L. T. Maloney, E. B. Johnston, and M. Young. Measurement and modeling of depth cue combination: in defense of weak fusion. *Vision Res*, 35(3):389–412, 1995.

J. Lewald, W. H. Ehrenstein, and R. Guski. Spatio-temporal constraints for auditory–visual integration. *Behav Brain Res*, 121(1-2):69–79, 2001.

W. J. Ma, J. M. Beck, P. E. Latham, and A. Pouget. Bayesian inference with probabilistic population codes. *Nat Neurosci*, 9:1432–8, 2006.

N. Macmillan and C. Crelman. *Detection theory: A users guide*. Lawrence Erlbaum Associates, 1991.

P. Mamassian and M. S. Landy. Interaction of visual prior constraints. *Vision Res*, 41(20): 2653–68, 2001.

D. W. Masaro. *Perceiving talking faces: from speech perception to a behavioral principle*. Cambridge, Massachusetts: MIT Press, 1998.

S. M. McClure, G. S. Berns, and P. R. Montague. Temporal prediction errors in a passive learning task activate human striatum. *Neuron*, 38(2):339–46, 2003.

W. McCulloch and W. Pitts. A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5:115–133, 1943.

J. O'Doherty, P. Dayan, K. Friston, H. Critchley, and R. J. Dolan. Temporal difference models and reward-related learning in the human brain. *Neuron*, 38(2):329–37, 2003.

J. O'Doherty, P. Dayan, J. Schultz, R. Deichmann, K. Friston, and R. J. Dolan. Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, 304(5669): 452–4, 2004. URL `http://dx.doi.org/10.1126/science.1094285`.

G. Orbán, J. Fiser, R. N. Aslin, and M. Lengyel. Bayesian model learning in human visual perception. In *Advances in Neural Information Processing*, volume 18, 2006.

M. G. Packard and B. J. Knowlton. Learning and memory functions of the Basal Ganglia. *Annu Rev Neurosci*, 25:563–93, 2002. URL `http://dx.doi.org/10.1146/annurev.neuro.25.112701.142937`.

J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference.* San Mateo, California: Morgan Kaufmann, 1988.

T. J. Perrault Jr, J. W. Vaughan, B. E. Stein, and M. T. Wallace. Neuron-specific response characteristics predict the magnitude of multisensory integration. *J Neurophysiol*, 90(6): 4022–6, 2003. URL `http://dx.doi.org/10.1152/jn.00494.2003`.

H. Pick, D. Warren, and J. Hay. Sensory conflict in judgements of spatial direction. *Percept. Psychophysics*, 6:203–205, 1969.

A. Pouget, S. Deneve, and J. R. Duhamel. A computational perspective on the neural basis of multisensory spatial representations. *Nat Rev Neurosci*, 3(9):741–7, 2002. URL `http://dx.doi.org/10.1038/nrn914`.

F.H. Previc. Functional specialization in the lower and upper visual fields in humans - its ecological origins and neurophysiological implications. *Behavioral and Brain Sciences*, 13:519–541, 1990.

R. Rao. Bayesian computation in recurrent neural circuits. *Neural Computation*, 16:1–38, 2004.

R. Rao, B. Olshausen, and M. Lewicki. *Probabilistic Models of the Brain*. Cambridge, Massachusetts: MIT Press, 2002.

J. H. Reynolds and R. Desimone. The role of neural mechanisms of attention in solving the binding problem. *Neuron*, 24(1):19–29, 111–25, 1999.

F. Rieke, D. Warland, R. de Ruyter van Steveninck, and W. Bialek. *Spikes: exploring the neural code*. MIT Press, Cambridge, MA, 1999.

D. D. Rife and J. Vanderkooy. Transfer-function measurements using maximum-length sequences. *J. Audio Eng. Soc*, 37(6):419–444, 1989.

N. W. Roach, J. Heron, and P. V. McGraw. Resolving multisensory conflict: a strategy for balancing the costs and benefits of audio-visual integration. *Proc Biol Sci*, 273(1598): 2159–68, 2006. URL `http://dx.doi.org/10.1098/rspb.2006.3578`.

K. S. Rockland and H. Ojima. Multisensory convergence in calcarine visual areas in macaque monkey. *Int J Psychophysiol*, 50(1-2):19–26, 2003.

A. L. Roskies. The binding problem. *Neuron*, 24(1):7–9, 111–25, 1999.

P. N. Sabes and M. I. Jordan. Reinforcement learning by probability matching. In Touretzky D. S., M. C. Mozer, and Hasselmo M. E., editors, *Advances in Neural Information Processing Systems*, volume 8, pages 1080–1086. The MIT Press, 1996. URL `citeseer.ist.psu.edu/sabes96reinforcement.html`.

R. Saxe, J. B. Tenenbaum, and S. Carey. Secret agents: inferences about hidden causes by 10- and 12-month-old infants. *Psychol Sci*, 16(12):995–1001, 2005. URL `http://dx.doi.org/10.1111/j.1467-9280.2005.01649.x`.

W. Schultz, P. Dayan, and P. R. Montague. A neural substrate of prediction and reward. *Science*, 275(5306):1593–9, 1997.

L. Shams, Y. Kamitani, and S. Shimojo. Illusions. What you see is what you hear. *Nature*, 408(6814):788, 2000. URL `http://dx.doi.org/10.1038/35048669`.

L. Shams, Y. Kamitani, and S. Shimojo. Visual illusion induced by sound. *Brain Res Cogn Brain Res*, 14(1):147–52, 2002.

L. Shams, W. J. Ma, and U. Beierholm. Sound-induced flash illusion as an optimal percept. *Neuroreport*, 16(17):1923–7, 2005.

D. Shanks, R. Tunney, and J. McCarthy. A re-examination of probability matching and rational choice. *Journal of Behavioral Decision Making*, 15:233250, 2002.

S. Shimojo, C. Simion, E. Shimojo, and C. Scheier. Gaze bias both reflects and influences preference. *Nat Neurosci*, 6(12):1317–22, 2003. URL `http://dx.doi.org/10.1038/nn1150`.

W. Singer. Neuronal synchrony: a versatile code for the definition of relations? *Neuron*, 24 (1):49–65, 111–25, 1999.

A. A. Stocker and E. P. Simoncelli. Noise characteristics and prior expectations in human visual speed perception. *Nat Neurosci*, 9(4):578–85, 2006. URL `http://dx.doi.org/10.1038/nn1669`.

R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

J. Tenenbaum. Bayesian modeling of human concept learning. In M. S. Kearns, S. A. Solla, and D. A. Cohn, editors, *Advances in Neural Information Processing*, volume 11, pages 59–65. Cambridge, MA: MIT Press, 1999.

G. Thomas. Experimental study of the influence of vision on sound localisation. *J Exp Psychol*, 28:167–177, 1941.

W. R. Thurlow and C. E. Jack. Certain determinants of the ventriloquism effect. *Percept Mot Skills*, 36(3):1171–84, 1973.

A. Treisman. The binding problem. *Curr Opin Neurobiol*, 6(2):171–8, 1996.

J. Trommershauser, L. T. Maloney, and M. S. Landy. Statistical decision theory and the selection of rapid, goal-directed movements. *J Opt Soc Am A Opt Image Sci Vis*, 20(7): 1419–33, 2003.

A. Tversky and D. Kahneman. Judgement under uncertainty: Heuristics and biases. *Science*, 185:1124–1131, 1974.

R. J. van Beers, A. C. Sittig, and J. J. Gon. Integration of proprioceptive and visual position-information: An experimentally supported model. *J Neurophysiol*, 81(3):1355–64, 1999.

H. von Helmholtz. *Handbuch der physiologischen Optic*. Hamburg, Germany: L. Voss, 1865-1867.

N. Vulkan. An economist's perspective on probability matching. *Journal of Economic Surveys*, 14:101–118, 2000. URL `citeseer.ist.psu.edu/vulkan98economists.html`.

M. R. Waldmann. Competition among causes but not effects in predictive and diagnostic learning. *J Exp Psychol Learn Mem Cogn*, 26(1):53–76, 2000.

M. T. Wallace, M. A. Meredith, and B. E. Stein. Integration of multiple sensory modalities in cat cortex. *Exp Brain Res*, 91(3):484–8, 1992.

M. T. Wallace, M. A. Meredith, and B. E. Stein. Multisensory integration in the superior colliculus of the alert cat. *J Neurophysiol*, 80(2):1006–10, 1998.

M. T. Wallace, G. E. Roberson, W. D. Hairston, B. E. Stein, J. W. Vaughan, and J. A. Schirillo. Unifying multisensory signals across time and space. *Exp Brain Res*, 158(2): 252–8, 2004. URL `http://dx.doi.org/10.1007/s00221-004-1899-9`.

Y. Weiss, E. P. Simoncelli, and E. H. Adelson. Motion illusions as optimal percepts. *Nat Neurosci*, 5(6):598–604, 2002. URL `http://dx.doi.org/10.1038/nn858`.

A. L Yuille and H. H. Bulthoff. Bayesian deicision theory and psychophysics. In *Perception as Bayesian Inference*. Cambridge University Press, 1996.