BRILL

Multisensory Research 26 (2013) 159–176

MULTISENSORY RESEARCH

brill.com/msr

Towards a Neural Implementation of Causal Inference in Cue Combination

Wei Ji Ma * **and Masih Rahmati**

Baylor College of Medicine, 1 Baylor Plaza, Houston TX 77030, Texas, USA

Received 3 June 2012; accepted 13 September 2012

Abstract

Causal inference in sensory cue combination is the process of determining whether multiple sensory cues have the same cause or different causes. Psychophysical evidence indicates that humans closely follow the predictions of a Bayesian causal inference model. Here, we explore how Bayesian causal inference could be implemented using probabilistic population coding and plausible neural operations, but conclude that the resulting architecture is unrealistic.

Keywords

Multisensory perception, cue combination, Bayesian inference, causal inference, modeling, population coding, neural networks

1. Introduction

The ventriloquist effect, whereby people misattribute a skilled performer's voice to a puppet (Howard and Templeton, 1966), is sometimes described as an illusion arising from near-optimal cue combination (Alais and Burr, 2004; Banks, 2004). The reasoning is that, because auditory localization is less precise (less reliable) than visual localization, the estimated location of origin of the speech will be closer to the visual than to the auditory event, leading the observer to attribute the speech to the puppet. If this cue integration explanation were correct, then the ventriloquist effect would occur even when the spatial disparity was large and the puppet's mouth movements would not match the performer's speech. However, in simplified experimental settings, ventriloquism breaks down at large spatial disparities (Slutsky and Recanzone, 2001; Wallace *et al.*, 2004), and experience with dubbed movies suggests that mis-

* To whom correspondence should be addressed. E-mail: wjma@bcm.edu

DOI:10.1163/22134808-00002407

matches in speech content reduce the illusion. In a more complete explanation of the ventriloquist effect, the observer first has to infer whether the auditory and the visual stimulus have a common cause and, only to the extent that they do, localize this cause at either the performer or the puppet. When disparity in space or speech content between auditory and visual signals is large, the observer will not believe that there is a common cause and simply perceives two separate events. Several years ago, two independent groups worked this idea out as a Bayesian model, called the causal inference model (Kording *et al.*, 2007; Sato *et al.*, 2007). In this model, which can be applied both to multisensory and to within-sensory cue combination, the observer computes the probability that the noisy measurements (cues) on a given trial were produced by the same cause (e.g. event location). The model provided a good fit to data from an experiment in which the subject reported whether a flash and a sound came from the same location or not (Kording *et al.*, 2007; Wallace *et al.*, 2004). In particular, the model quantitatively described how the proportion of 'common cause' reports depended on the spatial disparity between the stimuli. The causal inference model also predicted estimates of either the auditory or the visual stimulus location. When the probability of a common cause equals 1, this component of the model reduces to the well-known Bayesian model for cue *integration* (Clark and Yuille, 1990; Trommershauser *et al.*, 2011). Thus, the causal inference model provides a more complete account of multisensory phenomenology than the cue integration model. Causal inference likely also plays a role in scene segmentation, whether visual (Shams and Beierholm, 2010) or auditory (Bregman, 1990).

Although more behavioral experiments are needed to test the causal inference model, one theoretical challenge is clear: to determine how the computation of the probability of a common cause based on two noisy measurements can be implemented in neural circuitry. Less ambitiously, one might simply ask for a neural network to produce the same decision as the Bayesian model. However, it might be important for an organism to know not only whether a common cause was more likely than not, but also whether the probability of a common cause was, say, 51 or 99%. For example, if multiple pieces of information about the unity of the cause (say based on spatial disparity, based on temporal disparity, and based on speech content) had to be combined, then crucial information would be lost if each stream of information only outputted a 0 or a 1. Thus, we argue that not just the decision, but also the posterior probability of a common cause, must be accurately encoded.

Many proposals have been made for relating probability distributions to neural activity (Anastasio *et al.*, 2000; Anderson, 1994; Barlow, 1969; Berkes *et al.*, 2011; Deneve, 2008; Fiser *et al.*, 2010; Hoyer and Hyvarinen, 2003; Jazayeri and Movshon, 2006; Ma *et al.*, 2006; Pouget *et al.*, 2003; Rao, 2004; Shi *et al.*, 2010; Vilares and Kording, 2011; Zemel *et al.*, 1998). Here, we

use the framework that has been most successful so far in offering plausible, self-consistent neural implementations of Bayesian computations (Beck *et al.*, 2008, 2011; Ma *et al.*, 2006, 2011). In this framework, called probabilistic population coding (Ma *et al.*, 2006), a population of neurons encodes a likelihood function of a world state variable on every trial. We will first review the formalisms of Bayesian causal inference and probabilistic population coding. We will then use the latter to construct a neural implementation of the former.

2. The Causal Inference Model

The causal inference model is a Bayesian model of perception that applies when the observer receives multiple measurements that may or may not have the same cause. We restrict ourselves here to the case of two measurements, meaning that the number of possible causes is 1 or 2. An example would be to determine whether or not a synchronous flash and sound came from a common location. We first specify the statistical structure of the task, also called the generative model (Fig. 1). For simplicity, we assume that the probability that there is one cause in the world, $p(C = 1)$, equals 0.5, and so does, therefore, $p(C = 2)$. When $C = 1$, the common cause is a stimulus whose value is drawn from a stimulus distribution, $p(s)$. We assume this distribution is Gaussian with mean 0 and standard deviation σ_s, because that will make the calculations easier. (Another convenient choice could be the uniform distribution, were it not that this distribution cannot be normalized on the entire real line, and limiting it to an interval would make it much less convenient.) The stimulus produces two conditionally independent measurements, denoted x_1 and x_2, drawn from Gaussian distributions, both with mean s, but with potentially different standard deviations σ_1 and σ_2, respectively. When $C = 2$, there are two stimuli, s_1 and s_2, both drawn independently from the same distribution

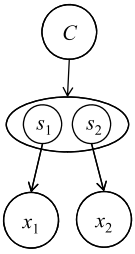

Figure 1. Generative model of causal inference. Nodes represent variables, arrows conditional dependencies. The common-cause variable C is of interest to the observer. When $C = 1$ (common cause), s_1 equals s_2. When $C = 2$ (different causes), s_1 and s_2 are independent. Independent Gaussian noise corrupts the scalar measurements x_1 and x_2. In the neural version of this model, the measurements are replaced by population patterns of activity, $\mathbf{r}_1$ and $\mathbf{r}_2$.

$p(s)$. Measurements x_1 and x_2 are then drawn from Gaussian distributions with means s_1 and s_2, and standard deviations σ_1 and σ_2, respectively.

The next step in our modeling process is to derive the observer's decision rule. On a given trial, the observer receives measurements x_1 and x_2 and is interested whether or not they have a common cause, that is, whether $C = 1$ or $C = 2$. The Bayesian observer makes this decision by calculating the probabilities of both hypotheses, $C = 1$ and $C = 2$, based on the measurements and perfect knowledge of the generative model. These probabilities sum to 1 and can therefore be characterized by a single number. It is convenient to express them as a log posterior ratio:

$$d = \log \frac{p(C = 1|x_1, x_2)}{p(C = 2|x_1, x_2)}. \tag{1}$$

Defining this ratio will simplify calculations and is convenient because the sign of d determines whether $C = 1$ or $C = 2$ is more probable. For example, if, on a given trial, $C = 2$ has twice the probability of $C = 1$, then $d = \log(0.5) = -0.69$. The inverse relationships are $p(C = 1|x_1, x_2) = \frac{1}{1+e^{-d}}$ and $p(C = 2|x_1, x_2) = \frac{1}{1+e^{d}}$. Applying Bayes' rule to equation (1), we find

$$d = \log \frac{p(x_1, x_2|C = 1)}{p(x_1, x_2|C = 2)} + \log \frac{p(C = 1)}{p(C = 2)} = \log \frac{p(x_1, x_2|C = 1)}{p(x_1, x_2|C = 2)}.$$

This is the logarithm of the ratio of the likelihoods of $C = 1$ and $C = 2$. We first consider $p(x_1, x_2|C = 1)$, the probability of the measurements x_1 and x_2 if there is a common cause. Since we do not know the stimulus s, we have to take into account every possible value of s. The probability of x_1 and x_2 given $C = 1$ in combination with a particular s would be $p(x_1, x_2|s)p(s|C = 1) = p(x_1, x_2|s)p(s)$. To find the total probability across all s, we integrate:

$$p(x_1, x_2|C = 1) = \int_{-\infty}^{\infty} p(x_1, x_2|s)p(s)\,\mathrm{d}s. \tag{2}$$

We now make use of the conditional independence of the measurements to write:

$$p(x_1, x_2|C = 1) = \int_{-\infty}^{\infty} p(x_1|s)p(x_2|s)p(s)\,\mathrm{d}s. \tag{3}$$

In this equation, $p(x_1|s)$ and $p(x_2|s)$ should be interpreted as functions of s: they express the sensory evidence on this trial and are called the likelihood functions of the stimulus. Thus, equation (3) expresses how the likelihood of $C = 1$ is computed from the likelihood functions of s.

We now turn to $p(x_1, x_2|C = 2)$. The logic is analogous, except that we have to integrate over two stimulus variables, s_1 and s_2:

$$\begin{aligned} & p(x_1, x_2|C = 2) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x_1, x_2|s_1, s_2)\, p(s_1, s_2|C = 2)\, \mathrm{d}s_1\, \mathrm{d}s_2 \\ &= \left(\int_{-\infty}^{\infty} p(x_1|s_1)\, p(s_1)\, \mathrm{d}s_1\right)\left(\int_{-\infty}^{\infty} p(x_2|s_2)\, p(s_2)\, \mathrm{d}s_2\right). \end{aligned} \tag{4}$$

As a result, the log posterior ratio is

$$d = \log \frac{\int_{-\infty}^{\infty} p(x_1|s)\, p(x_2|s)\, p(s|C = 1)\, \mathrm{d}s}{\left(\int_{-\infty}^{\infty} p(x_1|s_1)\, p(s_1)\, \mathrm{d}s_1\right)\left(\int_{-\infty}^{\infty} p(x_2|s_2)\, p(s_2)\, \mathrm{d}s_2\right)}. \tag{5}$$

After evaluating the integrals (see Appendix), this becomes

$$\begin{aligned} d = {} & \frac{J_1 J_2}{J_1 + J_2 + J_s}\left(x_1 x_2 - \frac{1}{2}\frac{J_1 x_1^2}{J_1 + J_s} - \frac{1}{2}\frac{J_2 x_2^2}{J_2 + J_s}\right) \\ & + \frac{1}{2}\log\left(1 + \frac{J_1 J_2}{J_s(J_1 + J_2 + J_s)}\right), \end{aligned} \tag{6}$$

where we have introduced the notation $J_1 = \frac{1}{\sigma_1^2}$, $J_2 = \frac{1}{\sigma_2^2}$, and $J_s = \frac{1}{\sigma_s^2}$ for the three precision (reliability) variables.

In Fig. 2(a), we plot the log likelihood ratio as a color code against the measurements, x_1 and x_2. The diagonal corresponds to trials on which the

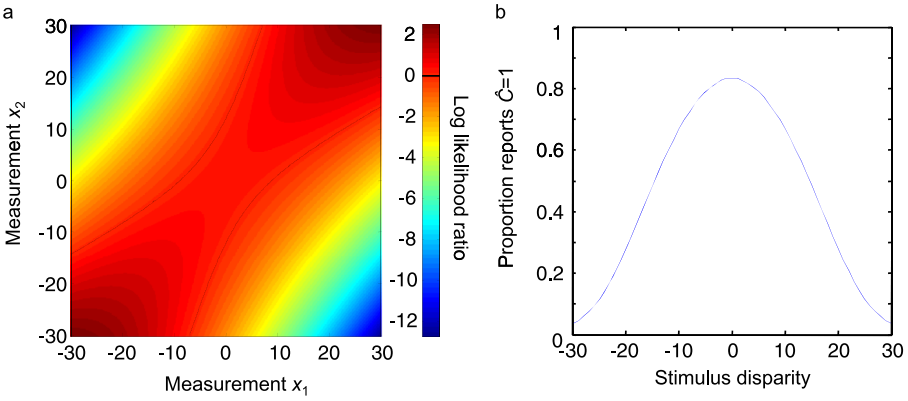

Figure 2. (a) The strength of the evidence in favor of a common cause, as expressed by the log likelihood ratio, as a function of the measurements x_1 and x_2. The $d = 0$ contour lines are shown in black. Two aspects of interest are the band around the diagonal and the structure within this band. Parameters were $\sigma_1 = 3$, $\sigma_2 = 10$, and $\sigma_s = 10$. (b) Proportion reports of a common cause as a function of stimulus disparity (s_2 minus s_1). This figure is published in colour in the online version.

measurements happen to be identical to each other. The hypothesis $C = 1$ becomes more likely relative to $C = 2$ when a pair of measurements lies closer to the diagonal. This is intuitive: when two measurements are similar, they are likely to have a common cause. In addition, we observe that the farther from 0 such a pair of similar measurements lies, the more likely they are to have a common cause. This is because we chose a stimulus distribution that peaks at 0. Even when there are two causes, the two stimuli and therefore the two measurements will tend to both lie near 0 and therefore close to each other. When measurements lie close to each other but far from 0, this is harder to explain as a consequence of the stimulus distributions, and it is therefore more likely that they have a common cause. (This also shows that the value of σ_s matters for the observer's decision.)

We model the final step in the decision process using a maximum-*a-posteriori* (MAP) rule: the observer chooses the hypothesis with the highest posterior probability. Applying the MAP decision rule maximizes expected accuracy and is in that sense optimal. In our task, the MAP observer reports that there was a common cause when d is positive. The observer's confidence in the decision can then be measured as the absolute value of d. Thus, in Fig. 2(a), the diagonal corners would correspond to high confidence in a 'common cause' decision ($\hat{C} = 1$), and the off-diagonal ones to high confidence in a 'different causes' decision ($\hat{C} = 2$).

Across many trials, we can compute the probability that the MAP observer reports 'common cause' ($\hat{C} = 1$) as a function of the true stimuli on a given trial, which we denote s_1 and s_2. This is equal to the probability that d is positive when x_1 and x_2 are generated by s_1 and s_2, respectively. We performed Monte Carlo simulation to calculate this probability. This entails randomly drawing pairs of x_1 and x_2 from their respective distributions $p(x_1|s_1)$ and $p(x_2|s_2)$ and counting for what proportion of these draws $d > 0$ holds. The resulting probability of reporting 'common cause' is plotted as a function of stimulus disparity, $s_2 - s_1$, in Fig. 2(b). We see that the larger the disparity between the two stimuli, the less frequently the observer reports that there is a common cause. This matches with empirical findings and the model also offers a good quantitative fit (Kording *et al.*, 2007; Sato *et al.*, 2007).

The causal inference model has not yet been tested in situations where the reliabilities of one or both cues are varied unpredictably from trial to trial, as is done in some cue integration experiments (Alais and Burr, 2004; Ernst and Banks, 2002; Landy and Kojima, 2001). Such unpredictable variations in reliability allow for a very strong test of the model, as they force the subject to take into account knowledge of σ_1 and σ_2 on every trial.

3. From a Behavioral to a Neural Model

The Bayesian model of causal inference outlined above is rather abstract: it largely regards the brain as a black box and simply represents observations as scalar variables following Gaussian distributions. While this model is adequate for describing behavior, it lacks mechanistic underpinnings. To find a neural implementation of the causal inference model, our starting point is that sensory information is represented in the brain through the spiking activity of populations of neurons. We then ask what kind of neural circuit can operate at the level of spike counts to produce behavior consistent with the Bayesian model and even compute, on each trial, the posterior probability of a common cause.

At the neural level, the observer's observations consist of a pattern of spike counts in a neural population, and the variability of this pattern across trials, for given stimuli, constitutes a generative model. We start by discussing a single population representing a stimulus. If $\mathbf{r}$ is the pattern of activity (a vector of spike counts, one for each neuron; Fig. 3), then $p(\mathbf{r}|s)$ quantifies the variability and is also called the noise distribution. For simplicity, we now assume that $\mathbf{r}$ is the activity in a population of independent Poisson neurons with Gaussian tuning curves; we will examine generalizations later. The tuning curve of the ith neuron is

$$f_i(s) = g A_i \mathrm{e}^{-\frac{(s-s_{\mathrm{pref},i})^2}{2\sigma_{\mathrm{tc},i}^2}}, \tag{7}$$

where g is a scaling factor, called gain, that is the same for all neurons in the population; A_i is this neuron's firing amplitude at unit gain; $s_{\mathrm{pref},i}$ is the neuron's preferred stimulus; and $\sigma_{\mathrm{tc},i}$ is the width of its tuning curve. We consider a general scenario in which stimulus reliability might change from trial to trial and affects neural population activity through gain, g (Ma *et al.*, 2006). We denote the distribution of gain by $p(g)$. The noise distribution of the population

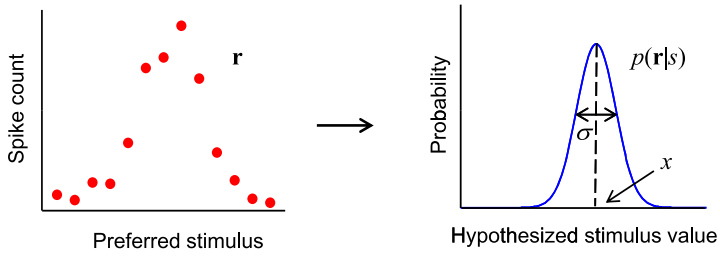

Figure 3. A population pattern of activity $\mathbf{r}$ encodes, on a single trial, a neural likelihood function of the stimulus. Note that although both plots have a roughly Gaussian shape, their interpretations are completely different and their widths will in general not be equal. This figure is published in colour in the online version.

for given g is:

$$p(\mathbf{r}|s,g) = \prod_{i=1}^{n} p(r_i|s) = \prod_{i=1}^{n} \frac{1}{r_i!} \mathrm{e}^{-f_i(s)} f_i(s)^{r_i}$$
$$= \left(\prod_{i=1}^{n} \frac{1}{r_i!} \right) \mathrm{e}^{-\sum_i f_i(s)} \mathrm{e}^{\sum_i r_i \log f_i(s)}. \quad (8)$$

We make the common assumption that $\sum_i f_i(s)$ is approximately independent of s, so that we can replace it by a constant K. We also substitute equation (7) for the tuning curves, to find:

$$p(\mathbf{r}|s,g) \approx K \left(\prod_{i=1}^{n} \frac{1}{r_i!} \right) \mathrm{e}^{\sum_i r_i \log g A_i} \mathrm{e}^{-\frac{1}{2}\sum_i r_i \frac{(s-s_{\mathrm{pref},i})^2}{\sigma_{\mathrm{tc},i}^2}}.$$

As in the behavioral model, the Bayesian observer (equations (3) and (4)) uses the likelihood function of s to compute the likelihood and posterior over C. The neural likelihood function of s is $p(\mathbf{r}|s)$ as a function of s (Fig. 3). Since gain is a random variable, this likelihood function is computed by marginalizing (averaging) $p(\mathbf{r}|s,g)$ over gain:

$$p(\mathbf{r}|s) = \int p(\mathbf{r}|s,g) p(g) \, \mathrm{d}g$$
$$\approx \int K \left(\prod_{i=1}^{n} \frac{1}{r_i!} \right) \mathrm{e}^{\sum_i r_i \log(g A_i)} \mathrm{e}^{-\frac{1}{2}\sum_i r_i \frac{(s-s_{\mathrm{pref},i})^2}{\sigma_{\mathrm{tc},i}^2}} p(g) \, \mathrm{d}g$$
$$= K \left(\prod_{i=1}^{n} \frac{1}{r_i!} \right) \mathrm{e}^{-\frac{1}{2}\sum_i r_i \frac{(s-s_{\mathrm{pref},i})^2}{\sigma_{\mathrm{tc},i}^2}} \int \mathrm{e}^{\sum_i r_i \log(g A_i)} p(g) \, \mathrm{d}g.$$

Importantly, the integral over g does not contain the stimulus and is therefore a constant in the likelihood function of s. Working out the factor containing s, we can see that the likelihood function is an unnormalized Gaussian. This Gaussian has a mean

$$x = \frac{\sum_i \frac{r_i s_{\mathrm{pref},i}}{\sigma_{\mathrm{tc},i}^2}}{\sum_i \frac{r_i}{\sigma_{\mathrm{tc},i}^2}} \equiv \frac{\mathbf{w}_{\mathrm{pref}} \cdot \mathbf{r}}{\mathbf{w}_{\mathrm{tc}} \cdot \mathbf{r}}, \quad (9)$$

which is also the maximum-likelihood estimate. (Equation (9) does not hold when the population is completely silent. Then, the likelihood function is completely flat.) The inverse variance of the normalized likelihood function (a measure of its width) is

$$\frac{1}{\sigma^2} = \sum_i \frac{r_i}{\sigma_{\mathrm{tc},i}^2} \equiv \mathbf{w}_{\mathrm{tc}} \cdot \mathbf{r}. \quad (10)$$

In equations (9) and (10), the weight vectors $\mathbf{w}_{\text{pref}}$ and $\mathbf{w}_{\text{tc}}$ are constants whose values are determined by the neurons' preferred stimuli and tuning widths. The inner product of these weight vectors with the spike counts of the neurons produces the maximum-likelihood estimate and likelihood width on a single trial. Biologically, these weights could be implemented as synaptic strengths. If the tuning curve width were independent of neuron, x would be the well-known center-of-mass or population vector decoder. The symbols x and σ have been re-introduced for a reason: the likelihood function of s is now

$$L(s) = p(\mathbf{r}|s) = \tilde{K}\,\mathrm{e}^{-\frac{(s-x)^2}{2\sigma^2}}, \tag{11}$$

which, up to the constant factor $\tilde{K}$ which contains all s-independent factors, is identical to the likelihood function of s in the behavioral model. In other words, the scalar measurements x_1 and x_2 that we used before are the values where the neural likelihood functions over s_1 and s_2, respectively, peak. We see in equation (10) that the inverse variance, $1/\sigma^2$, is encoded in the population as a weighted sum of the spike counts: this means that trial-to-trial variations in certainty could in principle be taken into account in downstream computation. Representing a neural likelihood function, such as the one in equation (11), in neural activity is known as probabilistic population coding (Ma *et al.*, 2006).

We can now reformulate the causal inference model in neural terms. The cues are represented in two populations, $\mathbf{r}_1$ and $\mathbf{r}_2$, each with their own $\tilde{K}$, x, and σ^2. The neural version of the log posterior ratio over C is:

$$d = \log \frac{\int_{-\infty}^{\infty} p(\mathbf{r}_1|s)\,p(\mathbf{r}_2|s)\,p(s|C=1)\,\mathrm{d}s}{(\int_{-\infty}^{\infty} p(\mathbf{r}_1|s_1)\,p(s_1)\,\mathrm{d}s_1)(\int_{-\infty}^{\infty} p(\mathbf{r}_2|s_2)\,p(s_2)\,\mathrm{d}s_2)}.$$

Substituting equation (11) for both populations, we see that the constant factors $\tilde{K}_1$ and $\tilde{K}_2$ drop out, implying that the distributions of the gains, $p(g_1)$ and $p(g_2)$, do not need to be known for the observer to be optimal. The log posterior ratio is identical to the one in the behavioral model, except that we have now made the identifications in equations (9) and (10). This means we can immediately jump to the final expression, equation (6), but now substitute equations (9) and (10), keeping in mind that $J_1 = \frac{1}{\sigma_1^2}$ and $J_2 = \frac{1}{\sigma_2^2}$:

$$d = \frac{(\mathbf{w}_{\text{pref},1}\cdot\mathbf{r}_1)(\mathbf{w}_{\text{pref},2}\cdot\mathbf{r}_2)}{\mathbf{w}_{\text{tc},1}\cdot\mathbf{r}_1 + \mathbf{w}_{\text{tc},2}\cdot\mathbf{r}_2 + J_s} - \frac{1}{2}\frac{(\mathbf{w}_{\text{tc},2}\cdot\mathbf{r}_2)(\mathbf{w}_{\text{pref},1}\cdot\mathbf{r}_1)^2}{(\mathbf{w}_{\text{tc},1}\cdot\mathbf{r}_1 + J_s)(\mathbf{w}_{\text{tc},1}\cdot\mathbf{r}_1 + \mathbf{w}_{\text{tc},2}\cdot\mathbf{r}_2 + J_s)}$$

$$-\frac{1}{2}\frac{(\mathbf{w}_{\mathrm{tc},1}\cdot\mathbf{r}_1)(\mathbf{w}_{\mathrm{pref},2}\cdot\mathbf{r}_2)^2}{(\mathbf{w}_{\mathrm{tc},2}\cdot\mathbf{r}_2+J_s)(\mathbf{w}_{\mathrm{tc},1}\cdot\mathbf{r}_1+\mathbf{w}_{\mathrm{tc},2}\cdot\mathbf{r}_2+J_s)}$$
$$+\frac{1}{2}\log\left(1+\frac{(\mathbf{w}_{\mathrm{tc},1}\cdot\mathbf{r}_1)(\mathbf{w}_{\mathrm{tc},2}\cdot\mathbf{r}_2)}{J_s(\mathbf{w}_{\mathrm{tc},1}\cdot\mathbf{r}_1+\mathbf{w}_{\mathrm{tc},2}\cdot\mathbf{r}_2+J_s)}\right), \tag{12}$$

where now all subscripts 1 and 2 refer to the two populations, not to individual neurons. This complicated expression is the optimal neural decision variable for causal inference when neural activity is independent Poisson and tuning curves are Gaussian. We have made only one approximation, namely that $\sum_i f_i(s)$ is independent of s; other than that, this result is exact.

Equation (12) also returns the correct answer when one or both populations are completely silent, namely $d = 0$.

In principle, one could implement causal inference with neurons by having neurons perform all operations in equation (12). The building blocks are four linear operations, so for convenience we let this preprocessing be done by four corresponding neurons:

$$\begin{aligned} z_{11} &= \mathbf{w}_{\mathrm{pref},1}\cdot\mathbf{r}_1, \quad & z_{12} &= \mathbf{w}_{\mathrm{tc},1}\cdot\mathbf{r}_1, \\ z_{21} &= \mathbf{w}_{\mathrm{pref},2}\cdot\mathbf{r}_2, \quad & z_{22} &= \mathbf{w}_{\mathrm{tc},2}\cdot\mathbf{r}_2. \end{aligned} \tag{13}$$

Then the log posterior ratio simplifies to

$$d = \frac{z_{11}z_{21}}{z_{12}+z_{22}+J_s} - \frac{1}{2}\frac{z_{22}z_{11}^2}{(z_{12}+J_s)(z_{12}+z_{22}+J_s)}$$
$$-\frac{1}{2}\frac{z_{12}z_{21}^2}{(z_{22}+J_s)(z_{12}+z_{22}+J_s)}$$
$$+\frac{1}{2}\log\left(1+\frac{z_{12}z_{22}}{J_s(z_{12}+z_{22}+J_s)}\right).$$

Each of the first three terms is a rational function of neural activities, with a polynomial of up to order 3 in the numerator and a polynomial of up to order 2 in the denominator. It is plausible that neurons can perform quadratic operations (Andersen *et al.*, 1985; Ben Hamed *et al.*, 2003; Boussaoud *et al.*, 1993; Bremmer *et al.*, 1997; Groh *et al.*, 2001; Trotter *et al.*, 1996) and therefore also multiplications. The division itself could be implemented using divisive normalization, which has been suggested to be widespread in cortex (Carandini and Heeger, 2011; Heeger, 1992). The fourth term is problematic, since it is not clear neurons can calculate logarithms, and a Taylor series expansion is not obviously meaningful.

This problem could be solved by approximating the fourth term by its trial average, so that the approximate decision variable would be equal to

$$d_{\text{approx}} = \frac{z_{11}z_{21}}{z_{12}+z_{22}+J_s} - \frac{1}{2}\frac{z_{22}z_{11}^2}{(z_{12}+J_s)(z_{12}+z_{22}+J_s)} - \frac{1}{2}\frac{z_{12}z_{21}^2}{(z_{22}+J_s)(z_{12}+z_{22}+J_s)} + \text{constant} \quad (14)$$

and the optimal decision rule would be $d_{\text{approx}} > 0$. To examine the goodness of this approximation, we performed a simulation.

4. Simulations

We simulated two input populations, each consisting of 100 independent Poisson neurons with Gaussian tuning curves, preferred stimuli $s_{\text{pref},i}$ randomly drawn from a uniform distribution on $[-70, 70]$, tuning curve widths $\sigma_{\text{tc},i}$ randomly drawn from a uniform distribution on [5, 35], and normalized amplitudes A_i randomly drawn from a uniform distribution on [0, 1]. A typical set of tuning curves is shown in Fig. 4(a). The probability of a common cause was 0.5. We chose the stimulus distribution to have a mean of 0 and a standard deviation of 10. We verified that the assumption that $\sum_i f_i(s)$ is approximately independent of s is satisfied over the relevant range $[-30, 30]$. Gain g was drawn independently for both populations from a gamma distribution with mean 1 and scale parameter 3. This means that the reliability of each stimulus varied unpredictably from trial to trial.

We simulated 100 000 trials. On each trial, we computed the log posterior ratio using equation (12). The histogram of each of the four terms in that expression separately across all trials is shown in Fig. 4(b). Visually, it appears that the fourth terms vary less than the other three. Indeed, averaged across 10 runs of 10^5 trials each, the standard deviations of the four terms were 5.22, 2.48. 3.63, and 0.37, respectively. This justifies approximating the fourth term by its trial average. This approximation leads to a decent approximation of the posterior probability of $C = 1$ (Fig. 4(c)). The 'distance' between the optimal posterior and the approximate posterior can be quantified by the Kullback–Leibler divergence (Cover and Thomas, 1991). We quantified information loss as the ratio of the trial-averaged Kullback–Leibler divergence to the mutual information between C and neural activity (Beck *et al.*, 2011; Ma *et al.*, 2011). For the parameters chosen and averaged over 20 runs, the information loss due to the approximation of the fourth term was 12%. Decisions based on d_{approx} coincided with decisions based on d on 91% of trials, and thus, the accuracy of the approximate observer was nearly the same as that of the optimal observer.

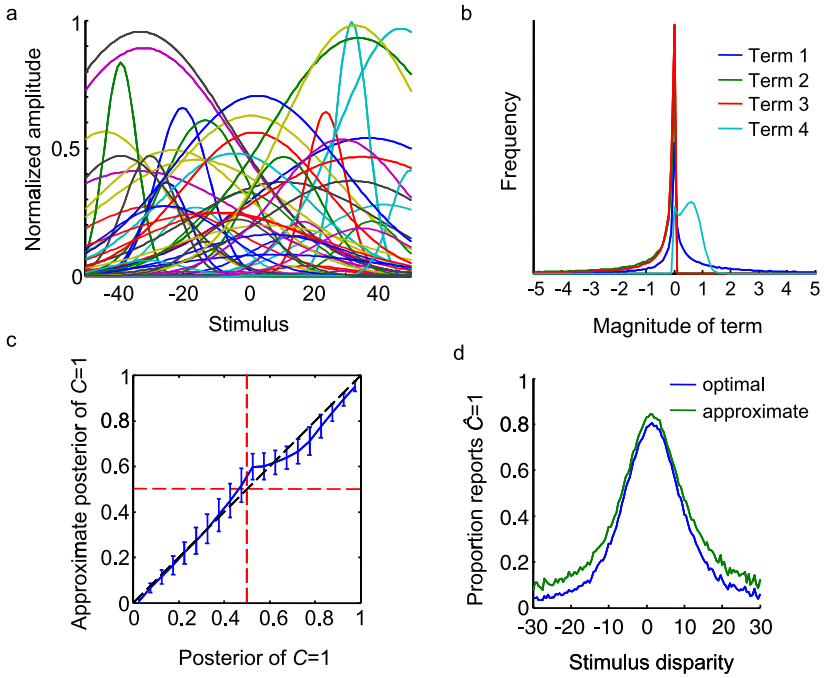

Figure 4. (a) Example set of tuning curves in an input population of sensory neurons used in the simulation. (b) Contribution of each of the four terms of the log likelihood ratio in equation (11). The fourth term has a much lower variance than the first three and we approximate it by its trial average. (c) Comparison of the posterior probability of a common cause estimated by an approximate network and the optimal posterior probability. Red lines indicate the decision criteria. Points in off-diagonal quadrants indicate deviations from the optimal observer. (d) Comparison of the proportion reports of a common cause as a function of stimulus disparity between the approximate network and the optimal observer.

By contrast, when we approximate any of the first three terms by its trial average, the approximation is extremely poor. For example, approximating the first term yields an information loss of 1600%, a coincidence rate of only 57%, and a drop in observer accuracy from 71 to 53%. Returning to the approximation of the fourth term, the probability of reporting a common cause is shown as a function of stimulus disparity in Fig. 4(d) for the optimal and approximate observer. As can also be seen in Fig. 4(c), the approximate observer reports, at every disparity, a common cause more often than the optimal observer. (Note that smaller disparities are more common than bigger ones.) All quantitative results reported here are specific to the parameters chosen, but the qualitative conclusions are robust under changes in parameters. We expect that in practice, when parameters have to be fitted to subject data, it is very difficult to distinguish between the optimal and the approximate strategy.

5. Generalizations

Some of our assumptions are easily generalized. Using a prior probability of a common cause different from 0.5 would simply introduce another constant term in the log posterior ratio. Another generalization is from independent Poisson variability to the exponential family with linear sufficient statistics (Ma *et al.*, 2006), which allows for Fano factors different from 1, continuous firing rates, tuning curves of arbitrary shapes, and correlated noise between neurons. The expression for such variability, also called Poisson-like variability, is

$$p(\mathbf{r}|s, g) = \varphi(\mathbf{r}, g)\mathrm{e}^{\mathbf{h}(s)\cdot\mathbf{r}}.$$

It is easily verified that equation (8) is a special case. For Poisson-like variability, like for independent Poisson variability with constant $\sum_i f_i(s)$, the distribution over gain g is irrelevant for the optimal decision rule. The final effect of generalizing to Poisson-like variability would only be that the linear weights in equation (13) would change.

6. Circuit

The computation in equation (14) can, in principle, be realized using the circuit diagram in Fig. 5, which contains three types of operations: linear combinations, quadratic nonlinearities and multiplications (counted as one type because $ab = ((a + b)^2 - a^2 - b^2)/2$), and divisive normalization. This results in a network that can, on each trial, reproduce, in good approximation, not only the optimal decision (the sign of d) but also decision confidence (the absolute value of d), even as the reliabilities of the two stimuli are unequal and vary unpredictably from trial to trial.

Although the network in Fig. 5 implements near-optimal causal inference, we mostly view it as an exposition of the limitations of a naïve probabilistic population coding approach than as a plausible circuit to look for in the brain. First, it is unclear how a complex network like this can be learned in limited time using biologically plausible learning rules. Second, in spite of the complexity of the network proposed here, it can only handle causal inference on two stimuli. If one were to take this approach for inferring whether a larger number of stimuli have a common cause (Wozny *et al.*, 2008), the number of operations needed would increase faster than linearly in the number of stimuli. Third, implementing a somewhat different but closely related task like same-different judgment, where in the $C = 2$ condition stimuli are drawn from a distribution around a common but trial-dependent mean (Van den Berg *et al.*, 2011), would ostensibly call for an entirely different circuit. Thus, our approach seems insufficiently general. Finally, causal inference does

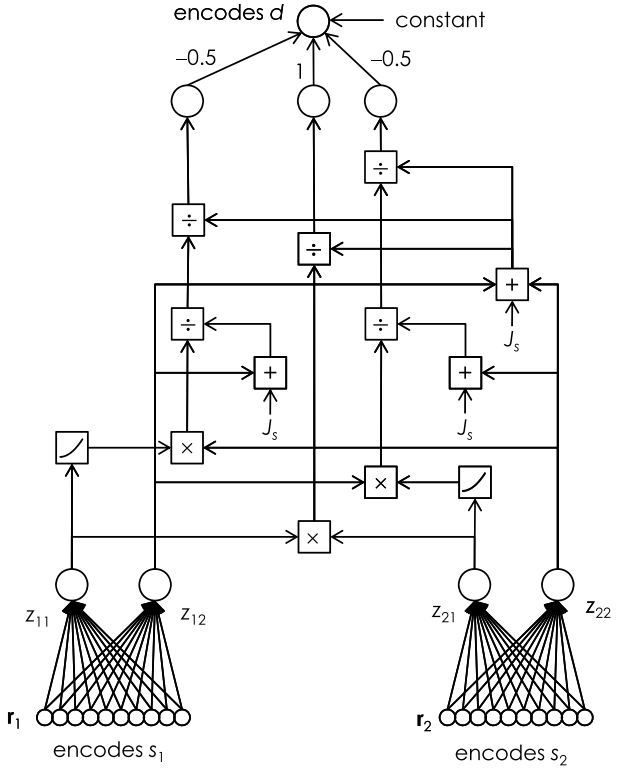

Figure 5. Circuit diagram of a network that can approximate the posterior probability of a common cause using linear combinations, quadratic operations, and divisive normalization. Input is assumed Poisson-like.

not only entail the computation of the probability of a common cause. In many cue combination tasks, a stimulus has to be estimated while the observer does not know whether or not there is a common cause (Kording *et al.*, 2007). This would add another layer of complexity beyond the computation of the posterior of C discussed here. To illustrate this: the mean of the posterior distribution of the first stimulus, s_1, is a weighted average of the posterior mean under the hypothesis that there is a common cause and the posterior mean under the hypothesis that there is not, with weights given by the posterior probabilities of $C = 1$ and $C = 2$:

$$\begin{aligned}\hat{s}_1 &= p(C=1|x_1,x_2)\hat{s}_{1,C=1} + p(C=2|x_1,x_2)\hat{s}_{1,C=2} \\ &= \frac{1}{1+\mathrm{e}^{-d}}\frac{\frac{x_1}{\sigma_1^2}+\frac{x_2}{\sigma_2^2}}{\frac{1}{\sigma_1^2}+\frac{1}{\sigma_2^2}+\frac{1}{\sigma_s^2}} + \frac{1}{1+\mathrm{e}^{d}}\frac{\frac{x_1}{\sigma_1^2}}{\frac{1}{\sigma_1^2}+\frac{1}{\sigma_s^2}}.\end{aligned}$$

Naively substituting equations (9), (10), and (14) would give rise to an unmanageably complex expression.

7. Discussion

The question how Bayesian computation is implemented in the brain is a central one in systems neuroscience. In previous work, we showed that optimal cue integration (with C always equal to 1) can, in a probabilistic population coding framework, be implemented using linear operations (Beck *et al.*, 2008; Ma *et al.*, 2006). The simplicity of these operations is one of the most appealing aspects of the framework. However, many generative models are much more complex than that of cue integration. A large part of this complexity is due to the presence of confounding or intermediate variables that need to be averaged out, such as the stimulus variables in equations (3) and (4). In earlier work, it was shown that this averaging process, called marginalization, can be implemented using a combination of three basic constituent elements: linear operations, quadratic operations, and divisive normalization (Beck *et al.*, 2011; Ma *et al.*, 2011). Here we have seen another example of such an implementation. Conceptually, this approach still very attractive, since prevalent operations in cortex, namely quadratic operations and divisive normalization, are linked to a prevalent operation in Bayesian inference, namely marginalization. However, for causal inference, we needed a large number of specific combinations of the constituent elements to realize the optimal decision rule, making the resulting architecture highly unrealistic. This is not due to excessive complexity of the generative model, but to our approach of literally translating the Bayesian decision rule, equation (6), to a neural rule using the 'dictionary' of equations (9) and (10).

A more promising direction, still using probabilistic population codes, might be to use variational methods (Bishop, 2006), to construct neural circuits that not only require fewer operations but are also more generally applicable. Instead of having one circuit for each task, it is likely that the brain employs general-purpose, heuristic inference machinery that can perform near-optimally in a large palette of related tasks, with only minor task-specific adjustments. Thus, the task-specific approach we took here might soon be superseded by a search for neural implementations of canonical approximate inference algorithms. Alternatively, frameworks other than probabilistic population codes (Anastasio *et al.*, 2000; Anderson, 1994; Barlow, 1969; Berkes *et al.*, 2011; Deneve, 2008; Fiser *et al.*, 2010; Hoyer and Hyvarinen, 2003; Jazayeri and Movshon, 2006; Rao, 2004; Shi *et al.*, 2010; Zemel *et al.*, 1998) might be able to provide biologically plausible approximations to optimal inference in tasks that require marginalization, but none have so far.

Acknowledgements

We thank Alex Pouget and Jeff Beck for numerous discussions.

References

Alais, D. and Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration, *Curr. Biol.* **14**, 257–262.

Anastasio, T. J., Patton, P. E. and Belkacem-Boussaid, K. (2000). Using Bayes' rule to model multisensory enhancement in the superior colliculus, *Neural Comput.* **12**, 1165–1187.

Andersen, R., Essick, G. and Siegel, R. (1985). Encoding of spatial location by posterior parietal neurons, *Science* **230**, 456–458.

Anderson, C. (1994). Neurobiological computational systems, in: *Computational Intelligence Imitating Life*, pp. 213–222. IEEE Press, New York, USA.

Banks, M. S. (2004). What you see and hear is what you get, *Curr. Biol.* **14**, R236–R238.

Barlow, H. B. (1969). Pattern recognition and the responses of sensory neurons, *Ann. N. Y. Acad. Sci.* **156**, 872–881.

Beck, J. M., Ma, W. J., Kiani, R., Hanks, T. D., Churchland, A. K., Roitman, J. D., Shadlen, M. N., Latham, P. E. and Pouget, A. (2008). Bayesian decision-making with probabilistic population codes, *Neuron* **60**, 1142–1145.

Beck, J. M., Latham, P. E. and Pouget, A. (2011). Marginalization in neural circuits with divisive normalization, *J. Neurosci.* **31**, 15310–15319.

Ben Hamed, S., Page, G., Duffy, C. and Pouget, A. (2003). MSTd neuronal basis functions for the population encoding of heading direction, *J. Neurophysiol.* **90**, 549–558.

Berkes, P., Orban, G., Lengyel, M. and Fiser, J. (2011). Spontaneous cortical activity reveals hallmarks of an optimal internel model of the environment, *Science* **331**, 83–87.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, Cambridge, UK.

Boussaoud, D., Barth, T. and Wise, S. (1993). Effects of gaze on apparent visual responses of frontal cortex neurons, *Exper. Brain Res.* **93**, 423–434.

Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press, Cambridge, MA, USA.

Bremmer, F., Ilg, U., Thiele, A., Distler, C. and Hoffman, K. (1997). Eye position effects in monkey cortex. I: Visual and pursuit-related activity in extrastriate areas MT and MST, *J. Neurophysiol.* **77**, 944–961.

Carandini, M. and Heeger, D. J. (2011). Normalization as a canonical neural computation, *Nat. Rev. Neurosci.* **13**, 51–62.

Clark, J. and Yuille, A. L. (1990). *Data Fusion for Sensory Information Processing Systems*. Kluwer, Norwell, MA, USA.

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. John Wiley and Sons, New York, USA.

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

Ernst, M. O. and Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion, *Nature* **415**, 429–433.

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

Groh, J. M., Trause, A. S., Underhill, A. M., Clark, K. R. and Inati, S. (2001). Eye position influences auditory responses in primate inferior colliculus, *Neuron* **29**, 509–518.

Heeger, D. J. (1992). Normalization of cell responses in cat striate cortex, *Vis. Neurosci.* **9**, 181–197.

Howard, I. P. and Templeton, W. B. (1966). *Human Spatial Orientation*. John Wiley and Sons, New York, USA.

Hoyer, P. O. and Hyvarinen, A. (2003). Interpreting neural response variability as Monte Carlo sampling of the posterior, in: *Neural Information Processing Systems*, Vol. 15. MIT Press, Cambridge, MA, USA.

Jazayeri, M. and Movshon, J. A. (2006). Optimal representation of sensory information by neural populations, *Nat. Neurosci.* **9**, 690–696.

Kording, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B. and Shams, L. (2007). Causal inference in multisensory perception, *PLoS ONE* **2**, e943.

Landy, M. S. and Kojima, H. (2001). Ideal cue combination for localizing texture-defined edges, *J. Optic. Soc. Amer. A* **18**, 2307–2320.

Ma, W. J. (2010). Signal detection theory, uncertainty, and Poisson-like population codes, *Vision Research* **50**, 2308–2319.

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

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

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

Rao, R. P. (2004). Bayesian computation in recurrent neural circuits, *Neural Comput.* **16**, 1–38.

Sato, Y., Toyoizumi, T. and Aihara, K. (2007). Bayesian inference explains perception of unity and ventriloquism aftereffect: identification of common sources of audiovisual stimuli, *Neural Comput.* **19**, 3335–3355.

Shams, L. and Beierholm, U. (2010). Causal inference in perception, *Trends Cognit. Sci.* **14**, 425–432.

Shi, L., Griffiths, T. L., Feldman, N. H. and Sanborn, A. N. (2010). Exemplar models as a mechanism for performing Bayesian inference, *Psychon. Bull. Rev.* **17**, 443–464.

Slutsky, D. A. and Recanzone, G. H. (2001). Temporal and spatial dependency of the ventriloquism effect, *Neuroreport* **12**, 7–10.

Trommershauser, J., Kording, K. and Landy, M. S. (Eds) (2011). *Sensory Cue Integration*. Oxford University Press, New York, USA.

Trotter, Y., Celebrini, S., Stricanne, B., Thorpe, S. and Imbert, M. (1996). Neural processing of stereopsis as a function of viewing distance in primate visual cortical area V1, *J. Neurophysiol.* **76**, 2872–2885.

Van den Berg, R., Vogel, M., Josic, K. and Ma, W. J. (2011). Optimal inference of sameness, *Proc. Nat. Acad. Sci. USA* **109**, 3178–3183.

Vilares, I. and Kording, K. P. (2011). Bayesian models: the structure of the world, uncertainty, behavior, and the brain, *Ann. New York Acad. Sci.* **1224**, 22–39.

Wallace, M. T., Roberson, G. E., Hairston, W. D., Stein, B. E., Vaughan, J. W. and Schirillo, J. A. (2004). Unifying multisensory signals across time and space, *Exper. Brain Res.* **158**, 252–258.

Wozny, D., Beierholm, U. and Shams, L. (2008). Human trimodal perception follows optimal statistical inference, *J. Vision* **8**(3), 1–11.

Zemel, R., Dayan, P. and Pouget, A. (1998). Probabilistic interpretation of population code, *Neural Computat.* **10**, 403–430.

Appendix

The Integral of the Product of Gaussians

Let $p_1(x), \ldots, p_N(x)$ be a set of N Gaussian probability density functions over x, with respective means $\mu_1, \ldots, \mu_N$, and standard deviations $\sigma_1, \ldots, \sigma_N$. The integral over the real line of the product of these Gaussians can be evaluated as

$$\begin{aligned}
&\int_{-\infty}^{\infty} \prod_{i=1}^{N} p_i(x)\,\mathrm{d}x \\
&\quad = \int_{-\infty}^{\infty} \prod_{i=1}^{N} \left(\frac{1}{\sqrt{2\pi\sigma_i^2}} \mathrm{e}^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}} \right) \mathrm{d}x \\
&\quad = \left(\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i^2}} \right) \sqrt{\frac{2\pi}{\sum_{i=1}^{N} \frac{1}{\sigma_i^2}}} \exp\left(-\frac{1}{2} \left(\sum_{i=1}^{N} \frac{\mu_i^2}{\sigma_i^2} - \frac{(\sum_{i=1}^{N} \frac{\mu_i}{\sigma_i^2})^2}{\sum_{i=1}^{N} \frac{1}{\sigma_i^2}} \right) \right).
\end{aligned}$$

We can apply this expression to numerator and denominator in equation (5). The numerator becomes

$$\begin{aligned}
&p(x_1, x_2|C=1) \\
&\quad = \frac{1}{2\pi} \sqrt{\frac{1}{\sigma_1^2\sigma_2^2\sigma_s^2}} \sqrt{\frac{1}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} + \frac{1}{\sigma_s^2}}} \\
&\qquad \times \exp\left(-\frac{1}{2} \left(\frac{x_1^2}{\sigma_1^2} + \frac{x_2^2}{\sigma_2^2} - \frac{(\frac{x_1}{\sigma_1^2} + \frac{x_2}{\sigma_2^2})^2}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} + \frac{1}{\sigma_s^2}} \right) \right).
\end{aligned}$$

The denominator becomes

$$\begin{aligned}
&p(x_1, x_2|C=2) \\
&\quad = \frac{1}{2\pi} \sqrt{\frac{1}{\sigma_1^2\sigma_2^2}} \frac{1}{\sigma_s^2} \sqrt{\frac{1}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_s^2}} \frac{1}{\frac{1}{\sigma_2^2} + \frac{1}{\sigma_s^2}}} \\
&\qquad \times \exp\left(-\frac{x_1^2}{2} \frac{\frac{1}{\sigma_1^2}\frac{1}{\sigma_s^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_s^2}} - \frac{x_2^2}{2} \frac{\frac{1}{\sigma_2^2}\frac{1}{\sigma_s^2}}{\frac{1}{\sigma_2^2} + \frac{1}{\sigma_s^2}} \right).
\end{aligned}$$

Substituting in equation (5) and simplifying yields equation (6).