Humans' multisensory perception, from integration to segregation, follows Bayesian inference

Ladan Shams[1,2], Ulrik Beierholm[3]

[1] *Department of Psychology, University of California, Los Angeles, Franz Hall 7445B, Los Angeles, California 90095-1563*

[2] *Division of Biology, California Institute of Technology, mail code 139-74, Pasadena, California, 91125*

[3] *Gatsby Computational Neuroscience Unit, UCL, London, UK*

Correspondence should be addressed to LS (ladan@psych.ucla.edu).

1 Introduction

Humans are almost always surrounded by multiple objects, and therefore multiple sources of sensory stimulation. 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 valid perception of the physical world, it must determine which of these temporally coincident sensory signals are caused by the same physical source, and thus should be integrated into a single percept. However, integration of sensory signals that originate from different objects/events can be detrimental, and therefore, it is equally important for the nervous system to determine which signals should be segregated. The signals of different modalities are more likely to be integrated if they are spatially congruent (Spence et al., 2004). For example, when in a parking lot, we tend to fuse the sound of a car honk with the image of a car which is spatially closest to the honk. Furthermore, signals are more likely to get integrated if they are structurally congruent. The sound of a TV program is normally integrated with the pictures on the TV screen, however, such integration would not occur if the speakers play the sound from a news program while the video displays a children's show. Thus, the determination of which set of temporally coincident sensory signals are to be bound together and which are to be segregated seems to be based on the degrees of spatial and structural consistency of the signals, which can vary from situation to situation.

In this chapter, we will first discuss experimental findings showing that multisensory perception encompasses a spectrum of phenomena ranging from full integration (or fusion), to partial integration, to complete segregation. Next, we will

describe two Bayesian causal inference models that can account for the entire range of combinations of two or more sensory cues. We will show that one of these models, which is a hierarchical Bayesian model, is a special form of the other one (which is a non-hierarchical model). We will then compare the predictions of these models with human data in multiple experiments, and show that Bayesian causal inference models can account for the human data remarkably well. Finally, we will present a study that investigates the stability of priors in the face of drastic change in sensory conditions.

2 Range of interactions between two modalities

We have explored the question of when and how the different sensory signals get integrated using two paradigms, one in which the degree of structural consistency between signals is varied, and one in which the degree of spatial consistency is varied. One paradigm uses the task of temporal numerosity judgment, and the other one the task of spatial localization. These two paradigms are complementary in the sense that the temporal numerosity judgment task primarily involves temporal processing whereas spatial localization is obviously a spatial task. In one case (numerosity judgment) hearing dominates the percept, whereas in the other (spatial localization) vision dominates under natural conditions. Another feature of these paradigms is that in both of them strong interactions occur between modalities under some conditions, leading to well-known illusions: the sound-induced flash illusion (Shams et al., 2000), and the ventriloquist effect (Howard and Templeton, 1966).

First, we discuss the human data in these two experimental settings. In the numerosity-judgment experiment (Shams et al., 2005), a variable number of flashes paired with a variable number of beeps (each ranging from zero to four) were presented in each trial and the subject was asked to report both the number of flashes and the

number of beeps they perceived at the end of each trial. The center of trains of flashes and beeps were temporally aligned in all trials. Fig. 1 shows the group data from 10 observers. Strong sound-induced flash illusions can be found in a large fraction of trials in the panels corresponding to 1flash+2beeps and 2flash+1beep where a single flash is perceived as two flashes, and where two flashes are perceived as one flash, respectively.

In the spatial localization experiment (Körding et al., 2007), observers were presented in each trial with either a brief visual stimulus in one of five locations (along the horizontal direction in the frontoparallel plane, ranging from -10° to +10° of visual field equally spaced), 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 by a key press in a forced-choice paradigm. All 35 combinations of these stimuli (except for no flash and no sound) were presented in pseudorandom order. The group data from 19 observers are shown in Fig. 2. The ventriloquist illusion can be found in several conditions, e.g., in V2A3 (i.e., visual stimulus at -5°, auditory at 0°) where the location of sound is captured by vision in a large fraction of trials.

In these illusion trials, in which perception in one modality is completely captured by the other modality or the percepts in two modalities are shifted towards each other and converge, the two modalities are fused. The two modalities are also fused when there is no discrepancy between the stimuli as in conditions along the diagonal in which the auditory and visual pulsations or locations are the same. Inspecting the matrix of joint auditory-visual responses (not shown here) confirms that indeed in these conditions the subjects generally report seeing and hearing the same thing simultaneously. At the other extreme, little or no interaction between the two modalities is found when there is a

large discrepancy between the two stimuli, e.g., in 1flash+4beeps or 4flash+1beep conditions (Fig. 1) or in V1A5 or V5A1 conditions (Fig. 2). More interestingly, there are also many trials in which the two modalities are neither completely fused nor segregated, but partially shifted towards each other. We refer to this phenomenon as "partial integration." It occurs when the discrepancy between the two modalities is moderate. Examples can be found in the 3flash+1beep condition in which two flashes and one beep are perceived (Fig. 1, 3), and in the V3A5 condition where V3 and A4 locations are perceived (Fig. 2) in a large fraction of trials. In summary, when the discrepancy between the two modalities is none or small, the two modalities get fused, when the discrepancy is large, the two modalities are segregated, and when the discrepancy is moderate, partial integration can occur (Fig 4a). Small discrepancy can be caused by environmental or neural noise even when the two signals originate from the same object. Therefore, it makes sense for the nervous system to fuse the signals when the discrepancy is minute. Similarly, large discrepancy is often due to the fact that the two signals are not caused by the same object, and therefore it makes sense for the nervous system to segregate them. It is not obvious, however, what the advantage of partial integration is for stimuli of moderate discrepancy. What would be gained by only partially integrating the information from two modalities? The goal of this work is to find computational principles that can explain this entire range of crossmodal interaction.

In order to examine whether there is a sharp divide between the situations of integration and segregation, we quantified the crossmodal integration in each condition (see Fig. 4 caption for details), and plotted the amount of integration as a function of discrepancy between the two modalities. Fig. 4b,c shows the results of this analysis for

the two experiments described above. As can be seen, integration between the two modalities degrades smoothly as the discrepancy between the two modalities increases. The model presented below accounts for this graded integration/segregation scheme using one single inference process.

It is not uncommon in daily life to have simultaneous stimulation in more than two sensory modalities. How do cues from three modalities interact? We tested trisensory cue combination using the numeriosity judgment paradigm where a variable number (ranging from 0 to 2) of flashes, beeps, and taps were presented in each trial and the subjects were asked to judge all three in each trial. Figure 5 shows the human data for the various stimulus conditions. This experiment revealed a number of different illusions, both of fusion (when two pulses are perceived as one) and fission (when one pulse is perceived as two) type, across almost all pairs and triplets of modalities. In almost all conditions where there was a discrepancy between the modalities, an illusion was found, showing that these interactions are indeed the rule rather than the exception in the human perceptual processing.

3 Bayesian model of integration and segregation

Traditional models of cue combination (see Fig. 6.a) (Alais and Burr, 2004; Bülthoff and Mallot, 1988; Knill, 2003; Landy et al., 1995; Yuille and Bülthoff, 1996) have been successful in accounting for the fusion phenomenon (e.g., (Alais and Burr, 2004; Ernst and Banks, 2002; Ghahramani et al., 1997; Hillis et al., 2002; van Beers et al., 1999)), but are unable to account for the segregation and partial integration range of interactions. As discussed above, the full integration only occurs when the discrepancy (spatial, temporal, structural) between the two signals is small. Moderate and large

discrepancies result in different percepts in different modalities, which cannot be explained by traditional models.

3.1 Implicit Causal Inference Model

We developed a Bayesian observer model (see Fig. 6.b) that does not assume such forced fusion (Shams et al., 2005). Instead, our model assumes a source for each signal; however, the sources are not taken to be statistically independent. Thus, the model allows inference about both cases in which separate entities have caused the sensory signals (segregation), and cases in which sensory signals are caused by one source (integration). The model uses Bayes' rule to make inference about the causes of the various sensory signals. This framework is very general and can be extended to any number of signals and any combination of modalities (see Fig. 6.c). However, for the purpose of illustration, we first focus on the combination of audition and vision.

We assume that the auditory and visual signals are statistically independent given the auditory and visual causes. This is a common assumption, motivated by the hypothesis that the noise processes that corrupt the signals in the different sensory pathways are independent. The information about the likelihood of sensory signal x_A occurring given an auditory cause s_A is represented by the probability distribution $P(x_A \mid s_A)$. Similarly, $P(x_V \mid s_V)$ represents the likelihood of sensory signal x_V given a source s_V in the physical world. The priors $P(s_A, s_V)$ denote the perceptual knowledge of the observer about the auditory-visual events in the environment. This prior probability encodes the interaction between the two modalities, and thus can be referred to as interaction priors. If the two modalities are independent of each other, the two dimensional distribution will be factorizable (e.g., like an isotropic Gaussian, see Fig.

7.a). In contrast, if the two modalities are expected to be consistent in all conditions, i.e., highly coupled together, then it will have a diagonal form (with all other non-diagonal values equal to zero, see Fig. 7.b). An ideal observer would try to make the best possible estimate of the physical sources s_A and s_V, based on the knowledge $P(x_A \mid s_A)$, $P(x_V \mid s_V)$ and $P(s_A, s_V)$. These estimates are based on the posterior probabilities $P(s_A, s_V \mid x_A, x_V)$, which can be calculated using Bayes' rule, and simplified by the assumptions represented by the model structure (Fig. 6.b), resulting in the following inference rule:

$$P(s_A, s_V \mid x_A, x_V) = \frac{P(x_A \mid s_A)\ P(x_V \mid s_V)\ P(s_A, s_V)}{P(x_A, x_V)} \quad . \qquad (1)$$

This inference rule simply states that the posterior probability of events s_A and s_V is the normalized product of the single-modality likelihoods and joint prior.

This model can account for the observer's data in both experiments (Beierholm, 2007; Shams et al., 2005). It can also be easily extended to combination of three modalities. Fig. 6.c shows the extension of the model to auditory-visual-tactile combination (Wozny et al., 2008). We tested this model on the trisensory numerosity judgement task. We assumed a univariate Gaussian distribution for likelihood functions, and a tri-variate Gaussian function for the joint prior. The standard deviations of the auditory, tatctile, and visual likelihood functions were estimated from unisensory conditions data. It was assumed that the means, variances and covariances of the multivariate prior distribution were all equal across the three senses, i.e., all three means are identical, all three variances are identical, and all three covariances are identical. Thus, the model had only three free parameters, corresponding to the variance, covariance and mean values of the prior distribution. To produce predictions of the

model, we ran Monte Carlo Simulations. On each trial, the mean of each likelihood function was assumed to be sampled from a Gaussian distribution with a mean at the veridical location and a standard deviation equal to that estimated from unisensory data. This results in different posterior distributions on different trials (of the same stimulus condition). We assume the observers minimize the mean squared error of their responses, and thus the optimal estimate would be the mean of the posterior distribution (which in this case is equivalent to the max, as the posterior is a guassian). In this fashion, for each stimulus condition, we obtain a distribution of responses, and this distribution was compared with the response distribution obtained from human observers. As can be seen in Fig. 5, the model can account for all two-way and three-way interactions. Using only three free parameters, the model can provide a remarkable account (R^2=0.95) for 676 data points.

3.2 Explicit Causal Inference Model

While the model described above can account for the entire range of interactions from integration, to partial integration, to segregation, it does not make explicit predictions about the perceived causal structure of the events. To be able to directly make predictions about the causal structure of the stimuli one needs a hierarchical model in which a variable encodes the choice between different causal structures (Fig. 8). Assuming that observers minimize the mean squared error of their responses, this generative model (Fig. 8) would result in the following inference. The probability of a common cause is determined using the Bayes rule as follows:

$$p\left(C=1 \mid x_V, x_A\right)=\frac{p\left(x_V, x_A \mid C=1\right) p\left(C=1\right)}{p\left(x_V, x_A\right)} \qquad (2)$$

In other words, the probability of common cause depends on the similarity between the two sensations and the prior belief in a common cause. More interestingly, the optimal estimate of the two modalities turns out to be a weighted average of two optimal estimates: the optimal estimate corresponding to independent-causes hypothesis, and the optimal estimate corresponding to the common-cause hypothesis, each weighted according to their respective probability:

$$\hat{s}=p\left(C=1 \mid x_V, x_A\right) \hat{s}_{C=1}+p\left(C=2 \mid x_V, x_A\right) \hat{s}_{C=2} \quad . \qquad (3)$$

This is a non-linear combination of the two sensations, and results in partial integration.

3.3 Relationship between the two models

Although the models shown in Eqs. 1 and 3 and Fig. 6.b and Fig. 8 look very different, they are intimately related. The hierarchical model is a special case of the non-hierarchical model. By integrating out variable C, the hierarchical model can be recast as a special form of the non-hierarchical model:

$$p(s_A, s_V \mid x_A, x_V)=\frac{p(x_A \mid s_A) p(x_V \mid s_V) p(s_A, s_V)}{p(x_A, x_V)}$$

where $p(s_A, s_V)=p(C=1) p(s)+p(C=2) p(s_A) p(s_V)$.

This formulation also makes it obvious that the hierarchical causal inference model is a mixture model, where the joint prior (Fig. 7.c) is a weighted average of the prior corresponding to common cause (Fig. 7.b) and a prior corresponding to the independent causes (an isotropic two-dimensional Gaussian, Fig. 7.a), each weighted by their respective prior probability. Mathematically, this model is equivalent to the mixture model proposed by Knill (Knill, 2003).

4 Comparison with Human Data

We compared the hierarchical-causal-inference model with human data in the two tasks discussed above; the spatial-localization task and the temporal-numerosity-judgment task. The prior and likelihoods were all modeled using Gaussian distributions. In the spatial localization task, mean of each likelihood function was sampled in each trial from a Gaussian distribution with a mean at the veridical location and a standard deviation that was a free parameter. The mean of the prior over space was at the center, representing a bias for straight-ahead location. As described earlier for the trisensory model, to produce model predictions we ran Monte Carlo simulations. We assume the observers minimize the mean squared error of their responses, and hence the optimal estimate would be the mean of the posterior distribution. In this fashion, for each stimulus condition, we obtain a distribution of responses, and this distribution was compared with the response distribution obtained from human observers. The model had four free parameters: the width of the visual likelihood, the width of the auditory likelihood, the width of the prior over space, and the prior probability of a common cause. These parameters were fitted to the data, and the results are shown in Fig. 2. As can be seen the model can account for the data in all conditions well. The model accounts for 97% of the variance in 1225 datapoints, using only 4 free parameters. The

same model can also account for the data reported by Wallace et al. (Wallace et al., 2004) where the subjects were asked to report their judgment of unity (common cause) (see Chapter XXX [chapter by Kording] for more detail).

Other models, including the traditional forced-fusion model, a model that does not integrate the stimuli at all, as well as two recent models that do not assume forced fusion were tested on the same dataset (data shown in Fig. 2, (Körding et al., 2007)), and compared with the causal-inference model. The causal-inference model outperformed all other models (Körding et al., 2007).

The hierarchical causal-inference model was also tested on the numerosity-judgment task. The data and model fits are shown in Fig. 1. As can be seen, the model also accounts for the data in all conditions well here. Accounting for 567 data points, the model explains 86% of the variance using only 4 free parameters.

In summary, the causal inference model can account for two complementary tasks, for data from two different laboratories, and outperforms other models. These results taken together suggest that human auditory-visual perception is consistent with Bayesian inference.

5 Independence of Priors and Likelihoods

The findings described above suggest that humans are Bayes-optimal in auditory-visual perceptual tasks given the (fitted) subjective priors. How is this achieved in the brain? The general understanding of Bayesian inference is that priors represent the statistics of the environment, whereas likelihoods

correspond to the sensory representations. In other words, it is generally assumed that likelihoods and priors are independent of each other. Does demonstration of Bayes-optimality under one condition indicate that the priors and likelihoods are independent? The answer to this question is no. The Bayes-optimality in a given task under a certain condition only implies that the performance of the observers is consistent with a Bayesian observer, but it could very well be the case that if we change the likelihoods, e.g. by modifying the stimuli, the priors would change and vice versa. Because we can estimate the likelihoods and priors using our model, we can empirically test this question.

Specifically, we asked whether priors are independent of likelihoods, i.e., whether a change in likelihoods would cause a change in priors. We tried to induce a change in visual likelihoods by altering the parameters of the visual stimuli. Namely, we tested two different visual stimulus contrasts in the same spatial localization task described earlier. To make sure that any potential learning of the priors within one session doesn't affect the results of the other session, we tested subjects in these two conditions one week apart, so that exposure to the statistics of real scenes would remove the effects of any possible learning within the first session. Therefore, the same observers were run in two sessions (low-visual-contrast and high-visual-contrast) using the same task, but with a one-week interval between the two sessions.

Indeed comparing the performance in the unisensory visual and auditory conditions between the two sessions, no significant difference was found in the auditory performance, whereas the visual performance was as much as 41% lower in the low-contrast conditions. The lack of change in the auditory performance is to be expected

because the auditory stimuli were not changed between the two sessions. The drastic change in visual performance confirms that the change in the visual stimulus was a substantial change. The model fits to the high-contrast condition were shown in Fig. 2. The model was also fitted to the low-contrast data and was found to account for the data well ($R^2 = .75$).

In summary, we find that the performance of the observers is Bayes-optimal in both high-contrast and low-contrast conditions given their subjective priors, and there is a substantial difference between the responses (and hence, the posteriors) in the low-contrast and high-contrast conditions. Therefore, it remains to be tested whether the likelihoods and priors are the same or different between the two sessions. If the priors are the same between the two sessions, then swapping the priors of the two sessions should not cause a change in the goodness of fit to the data. We tested this. We used the priors that were estimated from the low-contrast data to predict the high-contrast data, and vice versa. In both cases, the goodness of fit of the model remained mostly unchanged (going from $R^2 =0.97$ to $R^2 =0.95$ for the high contrast data, and from $R^2 =0.75$ to $R^2 =0.74$ for the low contrast data). This finding suggests that the priors are nearly constant between the two sessions.

We also compared the parameters of the likelihoods and prior distributions for the two sessions directly with each other. For the group data, the variance of auditory likelihood was highly similar between the two sessions, whereas the variance of visual likelihood was much larger for the low-contrast session. These results confirm that these parameters are not some arbitrary free parameters fitted to the data, but indeed capture the notion of likelihood functions. The prior parameters, namely the probability of a common cause and the width of the prior over space were highly similar between the

two sessions, suggesting that the priors were not different between the two sessions. In order to be able to test whether any of the differences between the parameters (albeit very small) are statistically significant, we fitted the model to the individual observers' data, and performed paired *t*-tests between the two sessions. The mean and standard errors of the four parameters across observers are shown in Fig. 9. As can be seen, there is no statistically significant difference for the prior parameters or auditory likelihood between the two conditions. The only parameter which is significantly different ($p<0.0005$) is the visual noise (i.e., standardard deviation of visual likelihood), which is much higher in the low-contrast condition. Therefore, these results indicate that despite a large change in likelihoods, the priors remained the same, suggesting that priors are indeed independent of likelihoods.

6 Discussion and Conclusions

The results presented here altogether suggest that the brain uses a mechanism similar to Bayesian inference for combining auditory, visual and tactile signals; for deciding whether or not, to what degree, and how (in which direction) to integrate the signals from two or more modalities. It should also be noted that two important and very different auditory-visual illusions (Howard and Templeton, 1966; Shams et al., 2000) can be viewed as the result of one coherent and statistically optimal computational strategy. We discussed a model that allows for both common cause and independent causes and can account for the entire range of interactions among auditory, visual and tactile modalities. This model only implicitly performs causal inference. We also presented a special form of this general model, a hierarchical model that explicitly assumes that one of two causal structures gave rise to the stimuli, and makes explicit predictions about the perceived causal structure of the stimuli. This model is useful for

accounting for data on judgments of unity, in experiments where the observers are explicitly probed for this judgment. It is however, not clear whether under natural conditions, observers do make a commitment to one or another causal structure. Future research can investigate this question empirically. While the hierarchical causal inference model has the advantage of making direct predictions about perceived causal structure of the events, it can become exceedingly complex for situations where three or more sensory signals are present simultaneously, as the number of possible causal structures becomes prohibitively large. For example, for the auditory-visual-tactile task discussed earlier, the non-hierarchical Bayesian model of Eq. 1 is substantially simpler than an extension of the hierarchical model of Eq. 3 to three modalities would have been.

While several previous studies have shown that human perception is consistent with a Bayesian observer (e.g., (Bloj et al., 1999; Stocker and Simoncelli, 2006; Weiss et al., 2002), the demonstration of Bayes-optimality does not indicate that the priors and likelihoods are encoded independently of each other as is the general interpretation of Bayesian inference. We discussed results that provide evidence for the priors being independent of likelihoods. This finding is consistent with the general notion of priors encoding the statistics of the environment and therefore being invariant to the sensory conditions at any given moment.

References

Alais, D., and Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. Curr Biol *14*, 257-262.
Beierholm, U. (2007). Bayesian modeling of sensory cue combinations. In Computation and Neural Systems (Pasadena, California Institute of Technology).
Bloj, M.G., Kersten, D., and Hurlbert, A.C. (1999). Perception of three-dimensional shape influences colour perception through mutual illumination. Nature *402*, 877-879.
Bülthoff, H.H., and Mallot, H.A. (1988). Integration of depth modules: stereo and shading. Journal of Optical Society of America *5*, 1749-1758.
Ernst, M.O., and Banks, M.S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. Nature *415*, 429-433.
Ghahramani, Z., Wolpert, D.M., and Jordan, M.I. (1997). Computational models of sensorimotor integration. In Self-organization, computational maps, and motor control, P.G. Morasso, and V. Sanguineti, eds. (Amsterdam, North-Holland: Elsevier Press), pp. 117-147.
Hillis, J.M., Ernst, M.O., Banks, M.S., and Landy, M.S. (2002). Combining sensory information: mandatory fusion within, but not between, senses. Science *298*, 1627-1630.
Howard, I.P., and Templeton, W.B. (1966). Human Spatial Orientation (London: Wiley).
Knill, D.C. (2003). Mixture models and the probabilistic structure of depth cues. Vision Research *43*, 831-854.
Körding, K., Beierholm, U., Ma, W.J., Tenebaum, J.M., Quartz, S., and Shams, L. (2007). Causal inference in multisensory perception. PLoS ONE *2*, e943.
Landy, M.S., Maloney, L.T., Johnston, E.B., and Young, M. (1995). Measurement and modeling of depth cue combination: In defense of weak fusion. Vision Res. *35*, 389-412.
Shams, L., Kamitani, Y., and Shimojo, S. (2000). What you see is what you hear. Nature *408*, 788.
Shams, L., Ma, W.J., and Beierholm, U. (2005). Sound-induced flash illusion as an optimal percept. Neuroreport *16*, 1923-1927.
Spence, C., Pavani, F., Maravita, A., and Holmes, N. (2004). Multisensory contributions to the 3-D representation of visuotactile peripersonal space in humans: evidence from the crossmodal congruency task. Journal of Physiology (Paris) *98*, 171-189.
Stocker, A.A., and Simoncelli, E.P. (2006). Noise characteristics and prior expectations in human visual speed perception. Nat Neurosci *9*, 578-585.
van Beers, R.J., Sittig, A.C., and Denier van der Gon, J.J. (1999). Integration of proprioceptive and visual position information: An experimentally supported model. J. Neurophysiol. *81*, 1355-1364.

Wallace, M.T., Roberson, G.H., Hairston, W.D., Stein, B.E., Vaughan, J.W., and Schirillo, J.A. (2004). Unifying mulitsensory signals across time and space. Exp Brain Res *158*, 252-258.
Weiss, Y., Simoncelli, E., and Adelson, E.H. (2002). Motion illusions as optimal percepts. Nature Neuroscience *5*, 508-510.
Wozny, D.R., Beierholm, U.R., and Shams, L. (2008). Human trimodal perception follows optimal statistical inference. Journal of Vision *8*, 24 21-11.
Yuille, A.L., and Bülthoff, H.H. (1996). Bayesian deicision theory and psychophysics. In Perception as Bayesian Inference, D.C. Knill, and W. Richards, eds. (Cambridge University Press), pp. 123-161.

Figure 1. Human observers response profiles in the temporal numerosity judgment task. 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 response distributions are displayed; i.e., in each condition, instead of showing 25 auditory-visual responses, 10 marginalized distributions are shown (5 auditory, and 5 visual). 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.

Figure 2. Human data in the spatial localization experiment. As in Fig. 1, to facilitate interpretation of the data, marginalized posteriors are shown. The representation of human are identical to those of Fig. 1, with the exception that the response categories are from left to right, -10 to +10 degrees along azimuth.

Figure 3. a) In the temporal numerosity judgment task, in the 3flash-1beep condition, one beep and two flashes are perceived in many trials (pointed by green arrow). b) This is an example of partial integration in which the visual percept is shifted towards the auditory percept but only partially.

Figure 4. a) The spectrum of perceptual phenomena as a function of degree of conflict between two modalities. When the discrepancy between two modalities is none or small, the signals tend to get fused. When the discrepancy is moderate, partial integration may occur. When the discrepancy is large, the signals are segregated. b) Bias (auditory influence on vision) as a function of discrepancy between the two modalities. Bias is here defined as the absolute deviation from veridical, divided by discrepancy. c) Bias (influence of vision on auditory perception) as a function of spatial disparity between the auditory and visual stimuli.

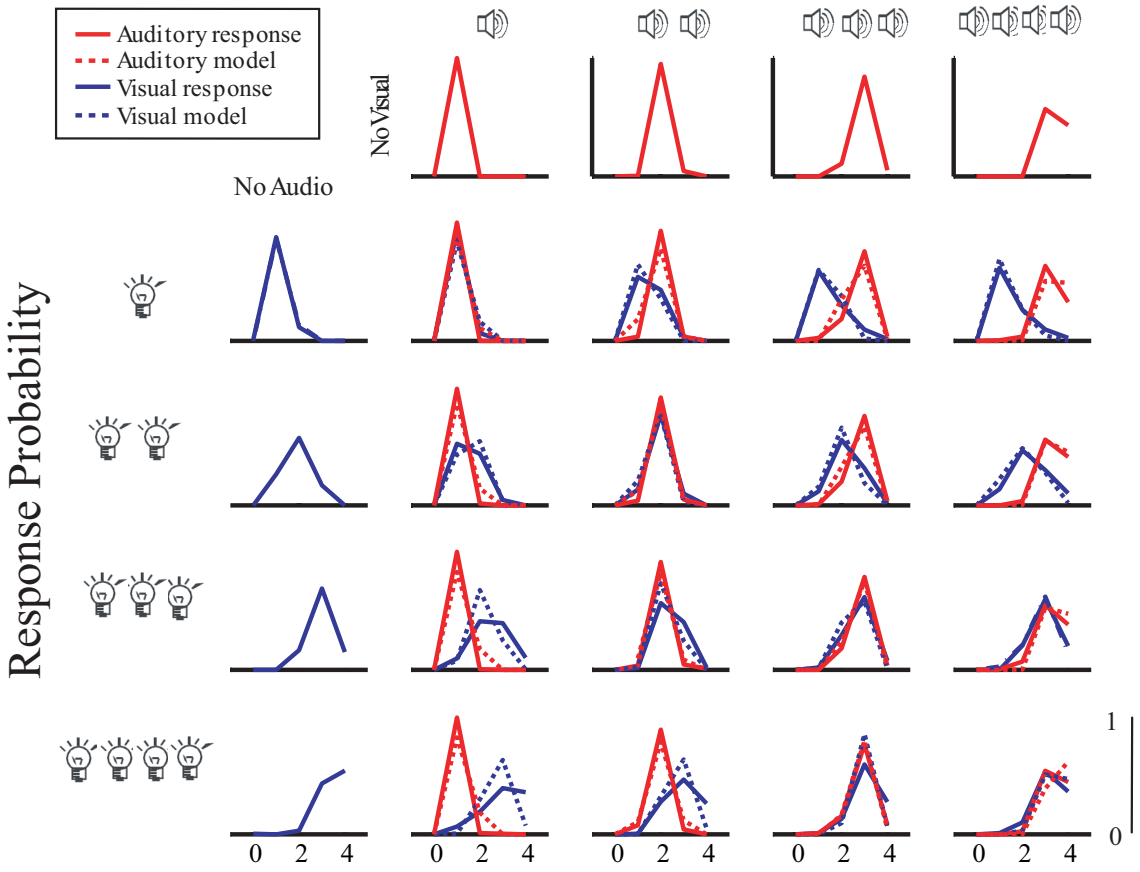
Figure 5. Human data and the Bayesian observer model fits in the tri-sensory numerosity judgment task. Each panel corresponds to one of the stimulus conditions. Blue, red, and black represent visual, auditory, and tactile responses, respectively. Symbols and solid lines represent data and broken lines represent model fits. Horizontal axis is the response category, and vertical axis the response probability.

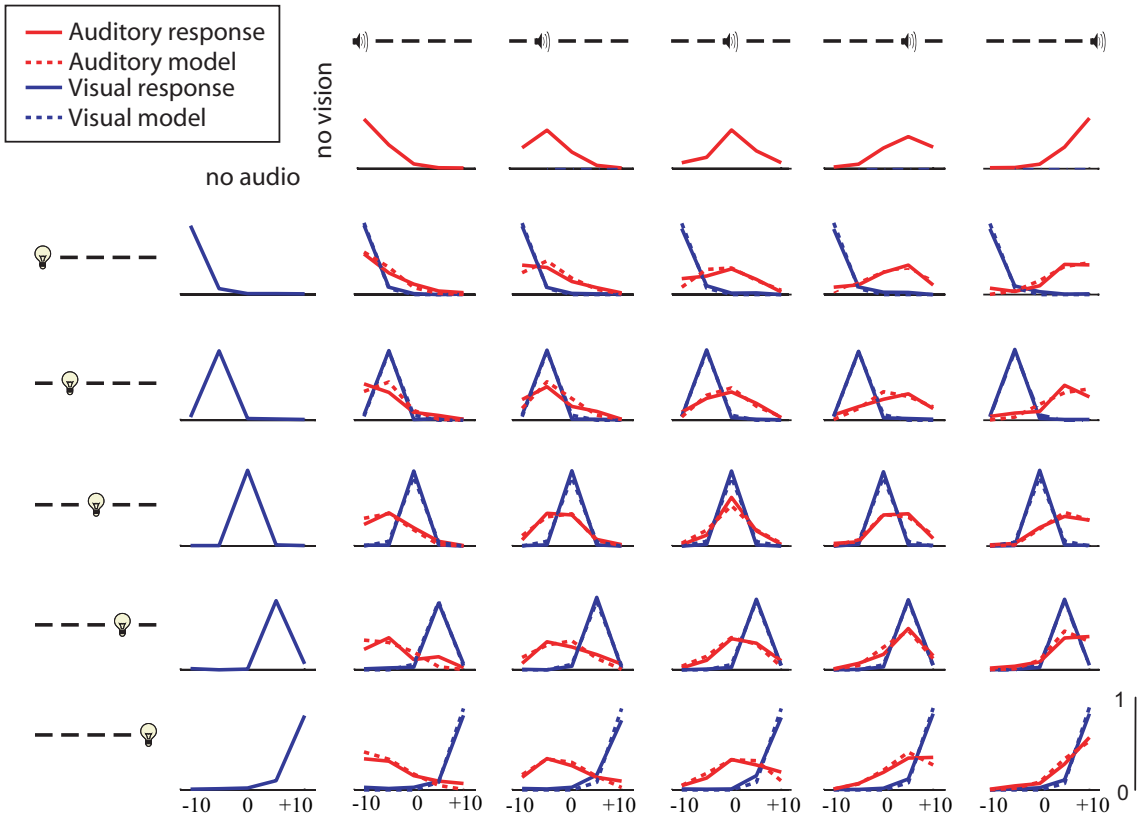
Figure 6. Generative models of different cue combination models. The graph nodes represent random variables, and arrows denote potential conditionality. The absence of an arrow represents statistical independence between the two variables. a) This Bayes net represents the traditional model of cue combination, in which the cues are assumed to be caused by one object. b) This graph represents the model of Shams et al. (2005) in which the two cues may or may not be caused by the same object. The double arrow represents a interdependency between the two variables. c) This graph represents the model of (Wozny et al., 2008) in which three cues are considered and any one or two or three of them may be caused by the same or distinct sources.

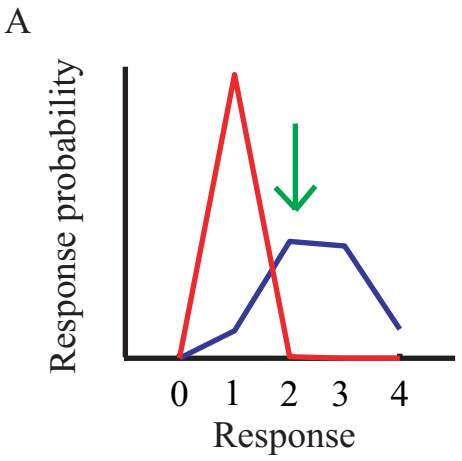
Figure 7. The interaction prior. The horizontal and vertical axes represent auditory and visual sources, respectively. A) This isotropic Gaussian function is factorizable and indicates that the two sources independent of each other. B) This prior indicates that the two sources are always the same, and cannot take on different values. C) This prior is a mixture of the priors shown in (a) and (b). This is the prior distribution for the causal inference model.

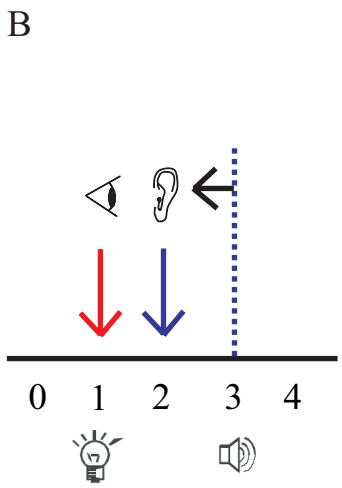
Figure 8. The generative model for the hierarchical causal inference model.

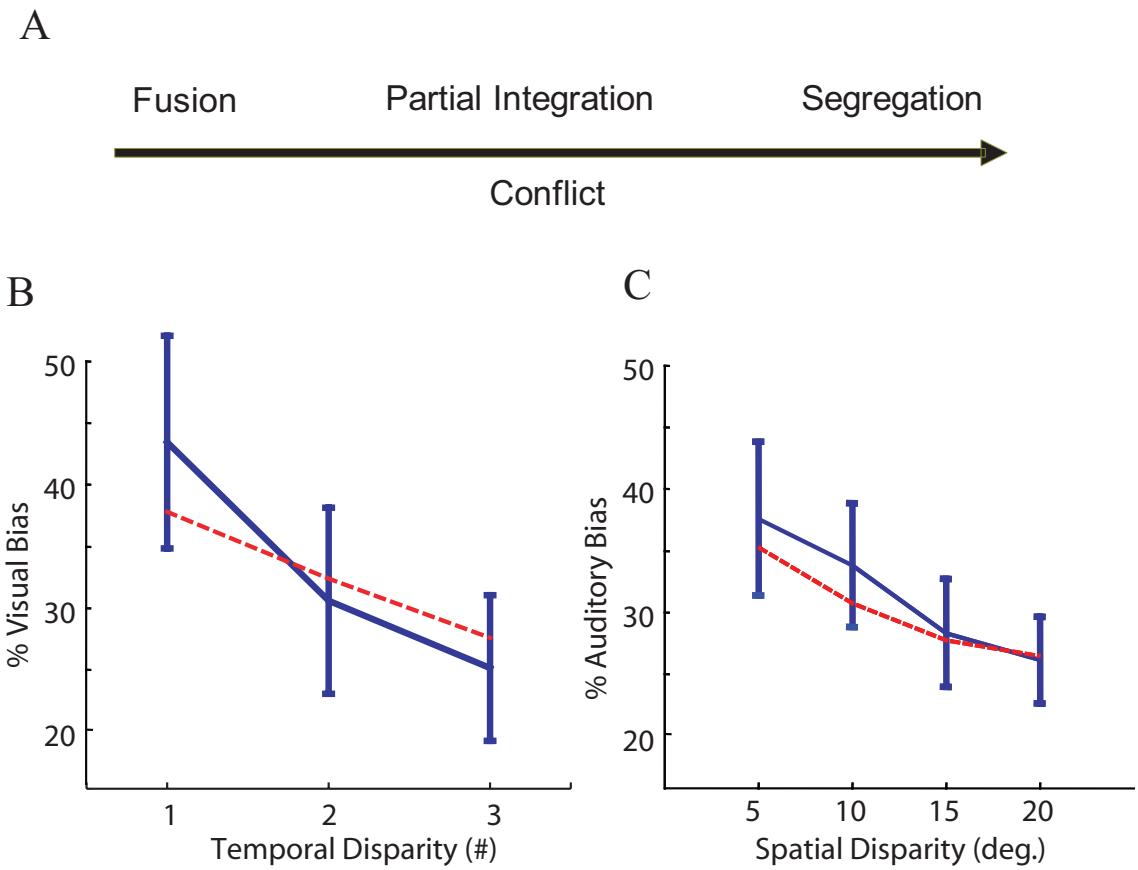
Figure 9. Mean parameter values across observers. Light grey bars represent values for the high-visual contrast condition, and dark grey bars represent those of low-contrast condition. Error bars denote the standard error of the mean.

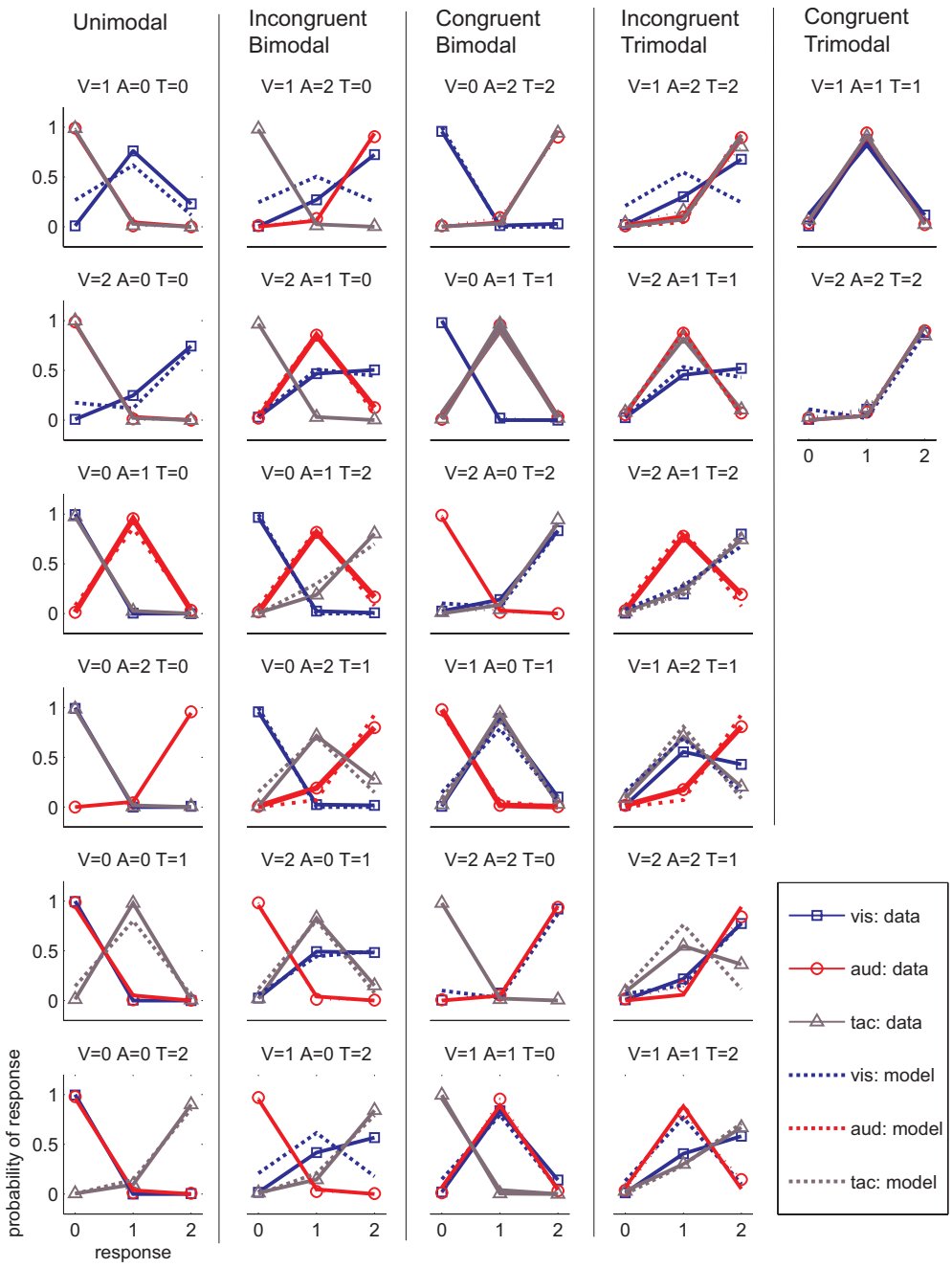


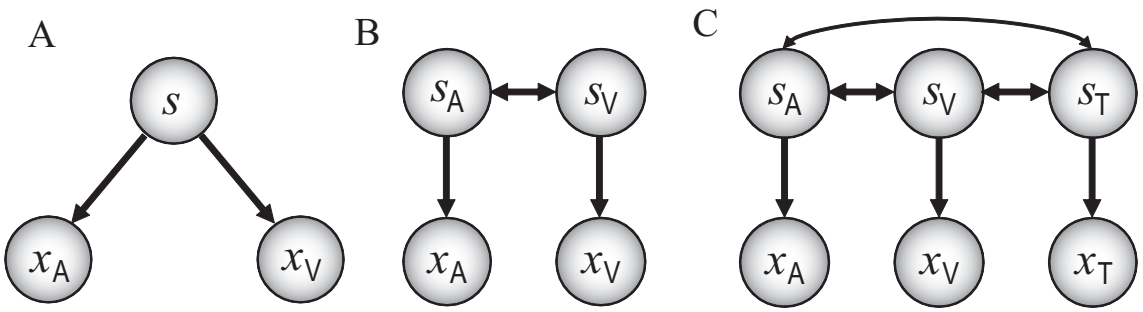

Auditory response
Auditory model
Visual response
Visual model
no vision
no audio
-10
0
+10
1
0

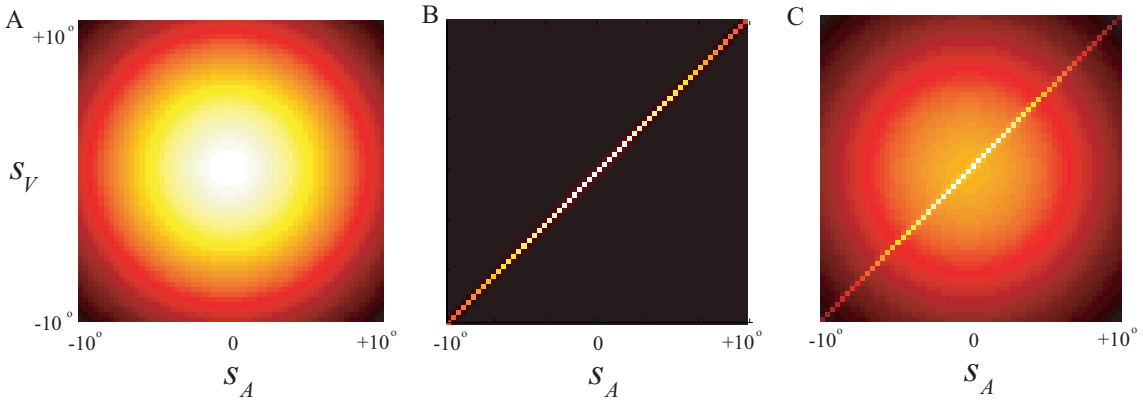

A
Response probability
0
1
2
3
4
Response


B
0
1
2
3
4


A
Fusion
Partial Integration
Segregation
Conflict
B
% Visual Bias
50
40
30
20
1
2
3
Temporal Disparity (#)
C
% Auditory Bias
50
40
30
20
5
10
15
20
Spatial Disparity (deg.)


Unimodal
Incongruent Bimodal
Congruent Bimodal
Incongruent Trimodal
Congruent Trimodal
V=1 A=0 T=0
V=1 A=2 T=0
V=0 A=2 T=2
V=1 A=2 T=2
V=1 A=1 T=1
V=2 A=0 T=0
V=2 A=1 T=0
V=0 A=1 T=1
V=2 A=1 T=1
V=2 A=2 T=2
V=0 A=1 T=0
V=0 A=1 T=2
V=2 A=0 T=2
V=2 A=1 T=2
V=0 A=2 T=0
V=0 A=2 T=1
V=1 A=0 T=1
V=1 A=2 T=1
V=0 A=0 T=1
V=2 A=0 T=1
V=2 A=2 T=0
V=2 A=2 T=1
V=0 A=0 T=2
V=1 A=0 T=2
V=1 A=1 T=0
V=1 A=1 T=2
probability of response
response
vis: data
aud: data
tac: data
vis: model
aud: model
tac: model

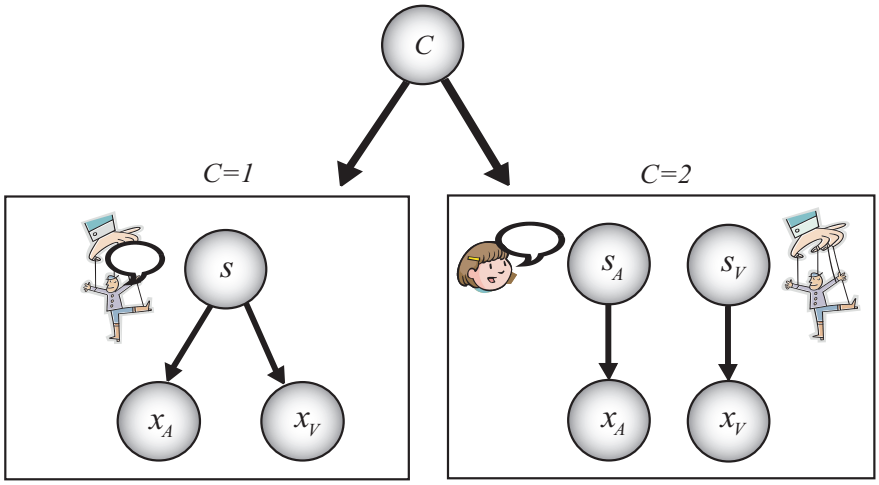


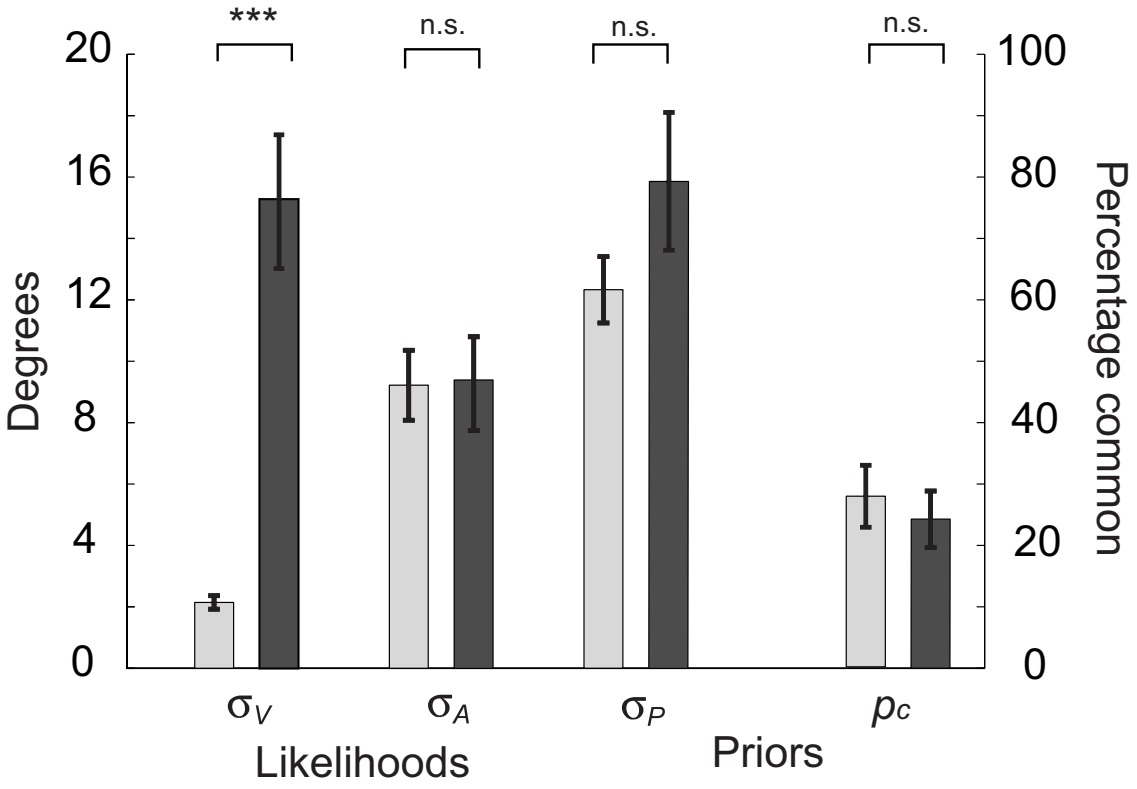


$p(s_V, s_A) = p(s_V)p(s_A)$

$p(s_V, s_A) = p(s)$

$p(s_A, s_V) = p(\text{common})p(s)$
$+p(\neg\text{common})p(s_A)p(s_V)$




***
n.s.
n.s.
n.s.
20
16
12
8
4
0
100
80
60
40
20
0
Degrees
Percentage common
σV
σA
σP
pc
Likelihoods
Priors