3 Probabilistic models for flexible model selection

3.1 Motivation

The Boltzmann machine represents generic priors, which can be seen as the superposition of many specific priors. However, there are also no mechanisms for partitioning the graph. Thus, in all situations, all the nodes (visual concepts) are relevant and used in global inference. We seek to develop a neural circuit mechanism that can bring out the appropriate prior model based on top-down task demands and bottom-up stimuli. This idea is related to dynamic functional circuitry, or the dynamic grouping of the relevant hypercolumns, to allow the formation of a cell assembly. Thus, this problem is one related to "model selection" in the Bayesian inference framework, coupled with graph partitioning.

3.2 Bayesian model selection

3.2.1 Problem setup

Each module can provide the neural representation for a number of visual attributes, such as luminance, edge orientation, direction of motion, stereo disparity, and color. For simplicity, we consider the orientation attribute in V1. In a neural implementation, this attribute could be represented in some embedded space of population neural activities, a probabilistic population code for example, where neurons may be discretized representatives of the visual attribute. For simplicity, we consider the inference of a particular attribute based on global information from multiple modules. Each module will encode an orientation attribute s_i with a set of orientation edge detectors, spanning the range of that attribute. This continuous variable can be considered as a continuum of visual concepts encoded in population activities of neurons, or it can be discretized as distinct "visual concepts".

The Boltzmann machine in the last section provides a formalism to learn the connections among the oriented detectors from the statistics of natural scenes. Here, we are interested in the flexible selection of these connections via binding or grouping of the modules during inference. We consider the basic mechanism for deciding what modules should be bound together for information integration during global inference. When two modules are coupled together, they are allowed to integrate their results and arrive at a common consensus of their input. For example, their combined responses can be used to determine the color, luminance, or texture of an object spanning a specific region of the image space. When they are not coupled together, they are assumed to have signals coming from distinct objects or regions. Thus, the grouping of modules is an example of model selection where different configurations of coupled modules can be considered as different models. In early vision, this is related to the problem of image segmentation ([18] for discussion). We have argued that such mechanisms can occur at every level of visual inference, and should be a universal mechanism (see [19] for example). In the early visual cortex, grouping and segmentation can take the form of condition or Markov random field and operate even at the image pixel level. This mechanism of grouping and segmentation can be generalized to many levels of visual reasoning. The modules could be encoding "visual concepts" using the embedded space or explicitly using prototype concept neurons. Thus, the grouping of modules can be considered as the binding of visual concepts. It does not have to be limited to elementary visual attributes. In higher areas of the visual cortex, grouping and segmentation operate at the level of object parts in the inference of global visual concept (i.e. objects). This can be implemented by feedforward, feedback, and/or horizontal recurrent connections. For the purpose of this report, we will use visual attributes and visual concepts interchangeably. In our model, we consider a visual concept to be a particular visual attribute, while visual concepts in

other settings (higher in the visual cortex or when we considered extreme sparse coding in V1) could be distinct and discrete prototypes.

In this section, we first show how the connection matrix, such as the one learned by the Boltzmann machine, can be decomposed into component connection matrices for each pair of modules. The "binding" between the two modules can be complete or partial. The degree of binding determines the degree of coupling of information (i.e. the posterior distribution) in the two modules during inference of the visual attributes in each module. It determines the degree of consensus of their shared beliefs. We set up a probabilistic model and analytically derive the mean and covariance of the posterior distributions. We show that they depend on the priors on the signal (x) covariance and the concept (s) covariance among the modules. We will finally show that the flexible binding of the modules (and the visual concepts encoded in them) can be formulated as a Bayesian model selection problem, with each model representing a possible configuration of module groupings. We propose that, in order to perform Bayesian model selection, the residual signals $D[s|x_i; s|x]$ of each module i relative to the consensus of the group will also need to be computed and maintained. The residual signal is the difference in the belief on the visual concept held by a module when it only consider its own direct bottom-up input versus when it considers all the input from its neighboring modules as well. This information is needed for computing the normalization factor during model section, as well as for recovering complete information during the synthesis process. Below are the detailed mathematical derivations on these points.

3.2.2 Likelihood function $p(\mathbf{x}|\mathbf{z})$

Suppose we have an observation $\mathbf{x} = (x_1, x_2, \ldots, x_n)^\top$, from n modules. Each x_i is independently generated by an underlying stimulus (visual concept) s_i in the ith module with following likelihood function

$$p(\mathbf{x}|\mathbf{s}) = \mathcal{N}[\mathbf{x}; \mathbf{s}, \Sigma_x], \tag{2}$$

where $\mathbf{s} = (s_1, s_2, \ldots, s_n)^\top$ is a vector of visual concepts (e.g. orientations) to be inferred at n modules, and the noise covariance matrix $\Sigma_x = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2)$ is a diagonal matrix because of conditional independence between x_i and x_j. The likelihood function can be expanded as a product over each x_i given s_i due to this conditional independence,

$$\begin{aligned} p(\mathbf{x}|\mathbf{s}) &= \prod_i p(x_i|s_i), \\ &= \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{(x_i - s_i)^2}{2\sigma_i^2}\right] \end{aligned} \tag{3}$$

3.2.3 Prior $p(\mathbf{s})$

We consider the prior of the visual concepts, $p(\mathbf{s})$, as a product of priors over pairs of s_i and s_j,

$$p(\mathbf{s}) \propto \prod_{i \neq j} p(s_i, s_j). \tag{4}$$

While the prior over each pair of $p(s_i, s_j)$ is defined as

$$p(s_i, s_j) = \frac{1}{L_s\sqrt{2\pi}\sigma_{ij}} \exp\left[-\frac{(s_i - s_j)^2}{2\sigma_{ij}^2}\right], \tag{5}$$

where L_s is the width of feature space. For example, $L_s = \pi$ for orientation. It is worthy to note that the marginal distribution $p(s_i) = \int p(s_i, s_j) ds_j = 1/L_s$ is a uniform distribution.

In $p(s_i, s_j)$, σ_{ij} effectively controls the correlation between s_i and s_j and thus determines the extent of integration between x_i and x_j in estimating s_i and s_j.

- $\sigma_{ij} = 0$ (full integration)
  In this case, the above Gaussian function collapses into a delta function $\delta(s_i - s_j)$, which indicates that s_i and s_j are exactly the same, and thus x_i and x_j will be integrated into a single quantity in the posterior [20].

- $\sigma_{ij} = \infty$ (segregation)
  In this case, s_i and s_j are independent, and the prior becomes $p(s_i, s_j) = 1/L_s^2$. x_i and x_j will be no long integrated. In other words, they will be segregated in estimating s_i and s_j.

- $0 < \sigma_{ij} < \infty$ (partial integration)
  In this case, s_i and s_j are correlated but not exactly the same. The estimate of either s_i will use the information from both x_i and x_j. A network implementation of this case corresponds to a decentralized decision architecture [21].

Next, we will formally define the integration and segregation models for s_i and s_j.

$$p(s_i, s_j|M) = \begin{cases} \frac{1}{\sqrt{2\pi} L_s \sigma_{ij}} \exp\left[-\frac{(s_i - s_j)^2}{2\sigma_{ij}^2}\right], & M = M_{int} \text{ (Integration, } \sigma_{ij} < \infty), \\ 1/L_s^2, & M = M_{seg} \text{ (Segregation, } \sigma_{ij} = \infty), \end{cases} \tag{6}$$

These forms of integration and segregation priors have been used in a number of psychophysical studies [21–24],

Matrix notation of prior $p(\mathbf{s})$

Now we consider a matrix notation of $p(\mathbf{s})$ to better understand its properties. In general, $p(\mathbf{s})$ can be denoted as

$$p(\mathbf{s}) \propto \exp\left[-\frac{1}{2}(\mathbf{s} - \mu_s)^\top \Sigma_s^{-1} (\mathbf{s} - \mu_s)\right]. \tag{7}$$

For the prior of two cues (Eq. 6), we have

$$\mu_s = (\mu, \mu)^\top, \tag{8}$$

$$\Sigma_s^{-1}|M_{int} = \sigma_{12}^{-2} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \tag{9}$$

$$\Sigma_s^{-1}|M_{seg} = 0 \times \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \tag{10}$$

It is worthy to note the following.

1. In the cases of integration or segregation, Σ_s^{-1} is a degenerate matrix because $det(\Sigma_s^{-1}) = 0$ and hence Σ_s^{-1} is non invertible. The degeneracy of Σ_s^{-1} will lead to the marginal prior $p(s_i)$ being a uniform distribution.

2. The mean of each s_i is the same, but the each s_i is uniformly distributed because Σ_s^{-1} is degenerate. Hence μ_s is like a free parameter.

Now we extend to the matrix notation for three modules, which can easily be generalized to n modules. The mean is

$$\mu_s = (\mu, \mu, \mu)^\top. \tag{11}$$

and the inverse of covariance matrix Σ_s^{-1} is

- When all three cues are integrated

$$\begin{aligned}
\Sigma_s^{-1} &= \sigma_{12}^{-2}\begin{pmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} + \sigma_{13}^{-2}\begin{pmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{pmatrix} + \sigma_{23}^{-2}\begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & -1 & 1 \end{pmatrix}, \\
&= \begin{pmatrix} \sigma_{12}^{-2}+\sigma_{13}^{-2} & -\sigma_{12}^{-2} & -\sigma_{13}^{-2} \\ -\sigma_{12}^{-2} & \sigma_{12}^{-2}+\sigma_{23}^{-2} & -\sigma_{23}^{-2} \\ -\sigma_{13}^{-2} & -\sigma_{23}^{-2} & \sigma_{13}^{-2}+\sigma_{23}^{-2} \end{pmatrix}
\end{aligned} \tag{12}$$

- When s_3 is segregated with s_2 ($\sigma_{23}^{-2} = 0$)

$$\begin{aligned}
\Sigma_s^{-1} &= \sigma_{12}^{-2}\begin{pmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} + \sigma_{13}^{-2}\begin{pmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{pmatrix} + 0\begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & -1 & 1 \end{pmatrix}, \\
&= \begin{pmatrix} \sigma_{12}^{-2}+\sigma_{13}^{-2} & -\sigma_{12}^{-2} & -\sigma_{13}^{-2} \\ -\sigma_{12}^{-2} & \sigma_{12}^{-2} & 0 \\ -\sigma_{13}^{-2} & 0 & \sigma_{13}^{-2} \end{pmatrix}
\end{aligned} \tag{13}$$

- When s_3 is segregated with both s_1 and s_2 ($\sigma_{13}^{-2} = 0$ and $\sigma_{23}^{-2} = 0$)

$$\begin{aligned}
\Sigma_s^{-1} &= \sigma_{12}^{-2}\begin{pmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} + 0\begin{pmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{pmatrix} + 0\begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & -1 & 1 \end{pmatrix}, \\
&= \begin{pmatrix} \sigma_{12}^{-2} & -\sigma_{12}^{-2} & 0 \\ -\sigma_{12}^{-2} & \sigma_{12}^{-2} & 0 \\ 0 & 0 & 0 \end{pmatrix}
\end{aligned} \tag{14}$$

We can see some properties of Σ_s^{-1} from above examples:

1. The diagonal element is the negative of the summation of the off-diagonal elements in the same row.
2. When s_i and s_j are segregated, $\sigma_{ij}^{-2} = 0$.
3. The Σ_s^{-1} matrix is the same as the effective connection matrix as the bump position dynamics (projected dynamics) of a continuous attractor neural network.

3.2.4 Inference of the posterior distributions

The inference of the posterior distribution $p(\mathbf{s}|\mathbf{x})$ is simple using matrix notation.

$$\begin{aligned} p(\mathbf{s}|\mathbf{x}) &\propto p(\mathbf{x}|\mathbf{s})p(\mathbf{s}), \\ &\propto \exp\left[-\frac{1}{2}(\mathbf{s}-\mathbf{x})^\top \Sigma_x^{-1}(\mathbf{s}-\mathbf{x}) - \frac{1}{2}(\mathbf{s}-\mu_s)^\top \Sigma_s^{-1}(\mathbf{s}-\mu_s)\right]. \end{aligned} \tag{15}$$

Since each marginal prior $p(s_i)$ is a uniform distribution[1] and each x_i is independently generated from s_i, the posterior of s_i can be solved as

$$\begin{aligned} p(s_i|\mathbf{x}) &= \int p(\mathbf{s}|\mathbf{x}) d\mathbf{s}_{\setminus i}, \\ &\propto \int \prod_{j=1}^{n} p(x_j|s_j) p(\mathbf{s}) d\mathbf{s}_{\setminus i}, \\ &\propto p(s_i|x_i) p(s_i| \cup_{j\neq i} x_j). \end{aligned} \tag{16}$$

In a special case that $\sigma_{ij} = 0$ and then $p(\mathbf{s}) = \prod_{i,j} \delta(s_i - s_j)$, above equation can be simplified into

$$p(s|\mathbf{x}) \propto \prod_{j=1}^{n} p(s|x_j), \ (i = 1, 2, \ldots, n), \tag{17}$$

where $s_1 = s_2 = \ldots = s_n \equiv s$. This equation is exactly the same as the posterior of information integration commonly seen in previous papers [20]. We see that the posterior $p(s|\mathbf{x})$ is a product of individual posteriors given each x_i, $p(s|x_i)$, which accumulate the information about s from each x_i.

With some algebra, the mean and covariance of $p(\mathbf{s}|\mathbf{x})$ can be derived as

$$Cov(\mathbf{s}|\mathbf{x})^{-1} = \Sigma_x^{-1} + \Sigma_s^{-1}, \tag{18}$$

$$\begin{aligned} \langle \mathbf{s}|\mathbf{x} \rangle &= Cov(\mathbf{s}|\mathbf{x})(\Sigma_x^{-1}\mathbf{x} + \Sigma_s^{-1}\mu_s), \\ &= Cov(\mathbf{s}|\mathbf{x})\Sigma_x^{-1}\mathbf{x}. \end{aligned} \tag{19}$$

In the last equality, $\Sigma_s^{-1}\mu_s = 0$ because Σ_s^{-1} is degenerate, and then the mean of prior disappears in the posterior like a free parameter. This free parameter μ_s is equivalent to the free parameter of bump position in continuous attractor.

$Cov(\mathbf{s}|\mathbf{x})$ is the covariance matrix of $\mathbf{s}$ conditioned on the stimulus $\mathbf{x}$, which can be regarded to the noise correlation matrix in physiology. $\mathbf{s}$ is a vector, with each entry corresponding to the chosen or inferred visual concept (orientation) in each module. The inference is about a visual concept or attribute s, which is to be represented by a population code of the neurons in a module, not by an individual neuron. Σ_x is the covariance matrix of the observation noise.

$\langle \mathbf{s}|\mathbf{x} \rangle$ denotes the mean of the posterior distribution of visual attributes across the modules given the input. Alternatively, the maximum a posterior estimate can also be computed.

3.2.5 Residual information is lost in the posterior and needs to be represented

As the network arrives at global consensus about a certain visual attribute, $\mathbf{s}$, using recurrent interaction via horizontal connections, the posterior of $\mathbf{s}$ at each module reflects the compromise

[1] A non-uniform Gaussian prior $p(s)$ can be considered as a likelihood from a constant input x_0.

of that module's opinions due to the opinions of the group. For decision on grouping and segmentation, it is important to keep the residual (difference) between each individual module's opinion and the global consensus. The strength of this "discomfort" with the global consensus will also ultimately drive the grouping and segmentation decision. A large residual will make a module to choose to be segregated from the group, and dissent from the global consensus. Also, keeping the residuals or the posterior of the residuals will allow the system to reconstruct the original input of the module by combining the global consensus and the residuals of the local modules.

Thus, apart from representing the posterior, we propose that each visual area needs to represent the residual information of input lost in the representation of the global consensus. Hence, we define the following residual function:

$$D[\mathbf{s}|x_i;\mathbf{s}|\mathbf{x}] = \frac{p(\mathbf{s}|x_i)}{[p(\mathbf{s}|\mathbf{x})]^{1/n}}, \tag{20}$$

where n is the number of x_i. Unlike only one posterior distribution $p(\mathbf{s}|\mathbf{x})$, there are n $\{D[s|x_i;s|\mathbf{x}]\}_{i=1}^{n}$. $D[\mathbf{s}|x_i;\mathbf{s}|\mathbf{x}]$ calculates the residual between the information of $\mathbf{s}$ conveyed by x_i only and that conveyed by all x_i. The $1/n$ exponent in the denominator ensures that the numerator and denominator have the same order.

When we have multiple modules, we can compute a general consensus $p(\mathbf{s}|\mathbf{x})$, which is an integrated, coherent interpretation of the scene. This $\langle\mathbf{s}\rangle$ can be represented by a set of congruent neurons, across the different modules, corresponding effectively to one global concept. In each module, this residual D is encoded using a set of opponent neurons in order to keep track of the difference between this module's own opinion before it listens to the group, and after it has arrives at a consensus with the group. The magnitude of this residual will contribute to the decision whether this module belongs to the selected group. If the residual is too large, it might choose not to belong to the party and becomes an independent.

We will illustrate this function with the implementation of the case of two modules (two input streams), Eq. (20) can be simplified into

$$\begin{aligned} D[\mathbf{s}|x_1;\mathbf{s}|\mathbf{x}] &= \frac{p(\mathbf{s}|x_1)}{p(\mathbf{s}|x_1,x_2)^{1/2}}, \\ &\propto \sqrt{\frac{p(\mathbf{s}|x_1)}{p(\mathbf{s}|x_2)}}, \end{aligned} \tag{21}$$

Compared with the posterior of $\mathbf{s}$ which takes the product of $\mathbf{s}$ given each module (Eq. 17), the residual information calculates the division of posterior $\mathbf{s}$ given each module by the consensus, which can be implemented by the opposite horizontal connections between opponent neurons in our network model, as will be shown in the below.

3.2.6 Bayesian model selection

Given the observations $\mathbf{x}$, the inference of $\mathbf{s}$ requires a system to choose a correct model (prior) to determine whether to include other x_j's to infer s_i. However, it remains unknown which prior should be used and which needs to be infered from input $\mathbf{x}$. In the case of two inputs x_1 and x_2, choosing one out of two priors needs us to compare their possibilities.

The posterior of the segregation model given the observation $\mathbf{x}$ can be calculated using Bayes

theorem. For example, the posterior of model 1 can be calculated to be,

$$\begin{aligned} p(M_1|\mathbf{x}) &= \frac{p(\mathbf{x}|M_1)p(M_1)}{p(\mathbf{x})}, \\ &= \frac{p(\mathbf{x}|M_1)p(M_1)}{\sum_j p(\mathbf{x}|M_j)p(M_j)}, \\ &= \frac{1}{1+\sum_j \frac{p(\mathbf{x}|M_j)}{p(\mathbf{x}|M_1)}\frac{p(M_j)}{p(M_1)}}, \end{aligned} \tag{22}$$

Depending on the number of modules involved, the number of groups or models could be quite large, each specifying a configurations of modules to be grouped or integrated during information sharing. For now, we focus on the 2 module case, which reduces to only two models, the segregation and integration models. For generalizing to a greater number of modules, we might have to explore mechanisms related to spectral graph theory (see Shi and Malik [25], Tolliver and Miller [26]).

3.2.7 Normalization Factor

The normalization factor is essential in Bayesian model selection. According to the definition of priors above, the covariance structure of prior $p(\mathbf{s})$ is given

$$\begin{aligned} p(\mathbf{x}|M) &= \int p(\mathbf{x}|\mathbf{s})p(\mathbf{s}|M)d\mathbf{s}, \\ &\propto \int \exp\left[-\frac{1}{2}(\mathbf{s}-\mathbf{x})^\top\Sigma_x^{-1}(\mathbf{s}-\mathbf{x}) - \frac{1}{2}(\mathbf{s}-\mu_s)^\top\Sigma_s^{-1}(\mathbf{s}-\mu_s)\right]d\mathbf{s}, \\ &\propto \exp\left[-\frac{1}{2}(\mathbf{x}-\mu_s)^\top\left(\Sigma_x^{-1}(\Sigma_x^{-1}+\Sigma_s^{-1})^{-1}\Sigma_s^{-1}\right)(\mathbf{x}-\mu_s)\right] \end{aligned} \tag{23}$$

Denote $\Sigma_{norm}^{-1} = \Sigma_x^{-1}(\Sigma_x^{-1}+\Sigma_s^{-1})^{-1}\Sigma_s^{-1}$. Note that $\left(\Sigma_x^{-1}(\Sigma_x^{-1}+\Sigma_s^{-1})^{-1}\Sigma_s^{-1}\right) \neq (\Sigma_x+\Sigma_s)^{-1}$ because Σ_s^{-1} is not invertible. We see Σ_{norm}^{-1} is also degenerate, because $det(\Sigma_{norm}^{-1}) = 0$.

M_j is a particular model that involves a particular configuration of modules. This gives us the probability of a particular model given the data in comparison to all other models (configurations). Note that this normalization factor is for normalizing the probability distribution. This is different from the divisive normalization in the neural circuits.

4 Network Circuit Model for Bayesian Model Selection

Here, we investigate a neural circuit model for implementing the probabilistic graphical model described in Section 3. The model currently assumes there are two modules, each module can be considered a hypercolumn that processes an input stream at a particular location. It can also be considered a network stream for processing a particular sensory cue, as in multi-cue integration. In our model, each module is populated with orientation selective neurons spanning 180 degrees. The probability model in Section 3 is general and can involve multiple modules. Here, for simplicity we consider the implementation of two modules and will extend this to include at least five modules as part of Phase 2. For our purpose, we can consider these two modules as either two adjacent hypercolumns or as a center hypercolumn, and the surround hypercolumn which represents the summary statistics of a set of surround hypercolumns. For most of the contextual modulation simulations, we assume the second scenario: a center and a surround hypercolumn.

The model has the following key components: (1) each module contains a population of neurons encoding a local visual concept, such as an orientation attribute using a population of orientation selective neurons in the form of probabilistic population code; (2) the neurons in the two modules are connected with excitatory connections to share information across each other for the inference of the global concept – these excitatory connections can be derived from stimulus co-occurrence statistics of the visual concepts (we assume neurons of similar orientation are more tightly coupled together, consistent with neural data and the Boltzmann machine results); (3) specific, orientation selective, feedforward, excitatory input projects to specific excitatory neurons (both the congruent and the opponent types), but the feedforward input also converges to the normalization neurons, which are not orientation selective; (4) the horizontal excitatory connections are gated by a model-selection, decision circuit that selects the prior models (when integration model is chosen, the gate is open, allowing the excitatory recurrent input from the surround to join the normalization of the neurons at the center, resulting in greater divisive normalization of the neurons that participate in a larger group than those participate in a smaller group); and (5) there are center-surround opponent neurons implemented by recurrent mutual excitation of neurons of opposite tunings, to track the residuals between the global consensus and the local opinions and the signals of these neurons will inhibit the integration model.

In the following section, we will provide details as to how the probability model is explicitly encoded in neuronal populations and implemented in network structures to perform grouping and segmentation in the framework of Bayesian model selection.

4.1 Probabilistic Model

4.1.1 Likelihood function $p(\mathbf{x}|\mathbf{s})$ and Prior $p(\mathbf{s})$

For two observations $\mathbf{x} = (x_1, x_2)^\top$ coming into the two modules, we assume each element x_i is *independently* generated by a latent variable s_i with the following likelihood function:

$$p(\mathbf{x}|\mathbf{s}) = \prod_{i=1}^{2} p(x_i|s_i). \tag{24}$$

The priors of $\mathbf{s}$ given the two models (integrated module M_{int} or segregated module M_{seg}) are given by:

$$p(s_1, s_2|M_{int}) = p(s_1, s_2), \tag{25}$$

$$p(s_1, s_2|M_{seg}) = p(s_1)p(s_2), \tag{26}$$

where $p(s_1) = \int p(s_1, s_2)ds_2$ in Eq. 26 and similar for $p(s_2)$. This implies that $p(s_1, s_2|M_{int})$ and $p(s_1, s_2|M_{seg})$ have same marginal priors $\{p(s_i)\}_{i=1,2}$. However, $p(s_1, s_2|M_{seg})$ deems $p(s_1)$ and $p(s_2)$ are independent and thus ignores the correlation information between them. The covariance information in the joint prior is the basis for information integration [21]. That is, the inference of s_1 not only uses the input from x_1, but also that from x_2 weighted by the covariance (see Section 3). Incidentally, the above priors for integration and segregation are also in accordance with those used in causal inference, e.g., [27, 28].

The likelihood function (Eq. 24) is typically modeled as a Gaussian distribution:

$$p(x_i|s_i) \quad = \quad \prod_{i=1}^{n} \mathcal{N}[x_i; s_i, \sigma_i^2], \tag{27}$$

where $\mathcal{N}[x_i; s_i, \sigma_i^2]$ is a univariate normal distribution over x_i with mean s_i and variance σ_i^2.

A convenient form of integration and segregation model priors used in psychophysical studies [21–24] is,

$$p(\mathbf{s}|M) = \begin{cases} \frac{1}{\sqrt{2\pi}L_s\sigma_s} \exp\left[-\frac{(s_1-s_2)^2}{2\sigma_s^2}\right], & M = M_{int} \text{ (Integration)}, \\ 1/L_s^2, & M = M_{seg} \text{ (Segregation)}, \end{cases} \tag{28}$$

where L_s is the width of feature space, e.g., 180° for orientation. It can be derived that the marginal prior of each s_i, e.g., $p(s_1) = \int p(s_1, s_2)ds_2$ is a *uniform* distribution. σ_s effectively controls the behavior of integration. When $\sigma_s = 0$, above Gaussian function collapses into a delta function $\delta(s_1 - s_2)$, which indicates that s_1 and s_2 are exactly the same and $\mathbf{x}$ will be integrated into a single quantity in posterior [20]. In another extreme that $\sigma_s = \infty$, meaning that $p(s_1)$ and $p(s_2)$ are independent, which corresponds to the segregation model considered in Eq. (26). When $0 < \sigma_s < \infty$, it means s_1 and s_2 are correlated but not exactly the same. A network implementation of this case corresponds to a decentralized architecture [21].

4.1.2 Reparameterization for circular variables

We assume the variable to be inferrred, $\mathbf{s}$, is the orientation of an edge. Because orientation is a circular variable, we model the likelihood and priors using *von Mises distributions*, which are variants of the circular Gaussian distribution that makes network implementation easier and analysis tractable. As we will see below, the Gaussian distribution can be seen as a local approximation of the von Mises distribution.

Suppose there are two visual concepts (e.g. orientation variables) s_1 and s_2 in the two modules, each of which generates a sensory observation x_m, for $m = 1, 2$ independently.

The likelihood function $p(x_m|s_m)$ von Mises distribution is

$$p(x_m|s_m) \quad = \quad \frac{1}{2\pi I_0(\kappa_m)} \exp\left[\kappa_m \cos(x_m - s_m)\right] \equiv \mathcal{M}(x_m - s_m, \kappa_m), \tag{29}$$

where $I_0(\kappa) = (2\pi)^{-1} \int_0^{2\pi} e^{\kappa\cos(\theta)} d\theta$ is the modified Bessel function of the first kind and order zero. s_m is the mean of the von Mises distribution, i.e., the mean value of x_m. κ_m is a positive number characterizing the concentration of the distribution, which is analogous to the inverse of the variance (σ^{-2}) of Gaussian distribution. In the limit of large κ_m, a von Mises distribution $\mathcal{M}[x_m - s_m, \kappa_m]$ approaches to a Gaussian distribution $\mathcal{N}[x_m - s_m, \kappa_m^{-1}]$. For small κ_m, the von Mises distribution deviates from the Gaussian one.

As an example in the present study, we choose the prior to be

$$p(s_1, s_2) = \frac{1}{2\pi}\mathcal{M}(s_1 - s_2, \kappa_s) = \frac{1}{(2\pi)^2 I_0(\kappa_s)} \exp\left[\kappa_s \cos(s_1 - s_2)\right]. \tag{30}$$

This form of prior favors the tendency for the two modules to share the similar visual concepts (e.g. similar orientations). κ_s determines the correlation between two visual concepts, i.e., how informative one observation is about the other, and it regulates the extent to which two observations in the two modules should be integrated. The fully integrated case, in which the prior becomes a delta function in the limit $\kappa_s \to \infty$, has been modeled in [20, 29]. It is worth noting that the marginal prior of either s_i is a uniform distribution, e.g., $p(s_1)$ can be calculated to be

$$p(s_1) = \int p(s_1, s_2) ds_2 = \frac{1}{2\pi}. \tag{31}$$

4.1.3 Posterior with information integration model

Now, we consider the interaction of the two modules when both belong to the same model, that is, are integrated together. Assuming that noises in the observations in the different modules are independent, the posterior distribution of two visual concepts can be written according to Bayes' theorem as

$$p(s_1, s_2|x_1, x_2) \propto p(x_1|s_1)p(x_2|s_2)p(s_1, s_2), \tag{32}$$

where $p(s_1, s_2)$ is the prior of the concepts, which specifies the concurrence probability and thus the interaction of a pair of visual concepts across modules – in this case, orientation stimuli at two separate locations.

Since the results for two visual concepts are exchangeable, we will henceforth only present the result for s_1, unless stated otherwise. Noting that $p(s_m) = p(x_m) = 1/2\pi$ are uniform distributions, (i.e. ignoring the cardinal effect in orientation) the posterior distribution of s_1 given observations in the two modules becomes

$$p(s_1|x_1, x_2) \propto p(x_1|s_1) \int p(x_2|s_2)p(s_2|s_1)ds_2 \propto p(s_1|x_1)p(s_1|x_2). \tag{33}$$

For module 1, the observation in module 2 x_2 is indirect, but it can still be informative to the inference of s_1 via the prior $p(s_1, s_2)$. By using Eqs. (29,30) and under reasonable approximations, we obtain

$$p(s_1|x_2) \propto \int p(x_2|s_2)p(s_2|s_1)ds_2 \simeq \mathcal{M}\left(s_1 - x_2, \kappa_{12}\right), \tag{34}$$

where $A(\kappa_{12}) = A(\kappa_2)A(\kappa_s)$ with $A(\kappa) \equiv \int_{-\pi}^{\pi} \cos(\theta) e^{\kappa \cos\theta} d\theta / \int_{-\pi}^{\pi} e^{\kappa \cos\theta} d\theta$.

Note that A is the mean resultant length in circular statistics, ranging from 0 to 1, where 0 means uniform distribution, and 1 means sharp distribution (i.e. delta function). These equations allow us to relate the different κ in the equations. κ_s is the concentration of the joint distribution between s_1 and s_2 or the inverse their covariance. κ_1 is the total synaptic input to all neurons in module 1, and κ_{12} is the total synaptic input to the neurons in module 1 from module 2, "filtered through" the connectivity matrix defined by the joint prior $p(s_1|x_1, x_2)$. In general, κ is the concentration of a von Mises distribution, corresponding the inverse of variance for a Gaussian distribution. It is related to reliability – a smaller variance means a higher concentration, reflecting "higher reliability" and confidence of either the consensus or discord. κ_s is related to σ_{12} in Equation 5. – when $\sigma_{12} = 0$, full integration takes place, when σ_{12} very large, segregation takes place.

Finally, utilizing Eqs. (29,34), Eq. (33) is written as

$$p(s_1|x_1, x_2) \quad \propto \quad \mathcal{M}(s_1 - x_1, \kappa_1)\mathcal{M}(s_1 - x_2, \kappa_{12}) = \mathcal{M}(s_1 - \hat{s}_1, \hat{\kappa}_1), \tag{35}$$

where the mean and concentration of the posterior given the observations in the two modules are:

$$\hat{s}_1 = \text{atan2}(\kappa_1 \sin x_1 + \kappa_{12} \sin x_2, \kappa_1 \cos x_1 + \kappa_{12} \cos x_2), \tag{36}$$

$$\hat{\kappa}_1 = \left[\kappa_1^2 + \kappa_{12}^2 + 2\kappa_1\kappa_{12}\cos(x_1 - x_2)\right]^{1/2}, \tag{37}$$

where atan2 is the arctangent function of two arguments. Note that $\hat{s}_1$ is the s_1 part of the mean estimate of the posterior given x_1, x_2, i.e. $\langle \mathbf{s}|\mathbf{x}\rangle$ and $\hat{\kappa}_1$ is the concentration, related to the module 1 part of the $Cov(\mathbf{s}|\mathbf{x})^{-1}$.

Eqs. (36,37) are the results of Bayesian integration in the form of von Mises distributions, and they are the criteria for judging whether optimal cue integration is achieved in a neural system.

To understand these optimality criteria intuitively, it is helpful to see their relation to the Gaussian distribution in the limit of large κ_1, κ_2 and κ_s. Under the condition $x_1 \approx x_2$, Eq. (37) is approximated to be $\hat{\kappa}_1 \approx \kappa_1 + \kappa_{12}$ (SI Sec. 2). Since $\kappa \approx 1/\sigma^2$ when von Mises distribution is approximated as Gaussian one, Eq. (37) becomes $1/\hat{\sigma}_1^2 \approx 1/\sigma_1^2 + 1/\sigma_{12}^2$, which is the Bayesian prediction on Gaussian variance conventionally used in the literature [20]. Similarly, Eq. (36) is associated with the Bayesian prediction on the Gaussian mean [20].

4.1.4 Residual information

According to the definition, the residual information between the observations from the two channels $D(s_1|x_1; s_1|x_2)$ is measured by the ratio of the posteriors given each observation,

$$D(s_1|x_1; s_1|x_2) \equiv p(s_1|x_1)/p(s_1|x_2), \tag{38}$$

By taking the expectation of $\log D$ over the distribution $p(s_1|x_1)$, we get the Kullback-Leibler divergence between the two posteriors given each observation.

Interestingly, by utilizing the properties of von Mises distributions and the condition $\cos(s_1 + \pi - x_2) = -\cos(s_1 - x_2)$, Eq. (38) can be rewritten as

$$D(s_1|x_1; s_1|x_2) \propto p(s_1|x_1)p(s_1 + \pi|x_2). \tag{39}$$

Intuitively, the residual information between the observations from in the two modules is proportional to the product of the posterior given the direct observation and the posterior given the indirect observation in another module but with the attribute orientation value shifted by π in this case.

By utilizing Eqs. (29,34), we obtain

$$D(s_1|x_1; s_1|x_2) \propto \mathcal{M}(s_1 - x_1, \kappa_1)\mathcal{M}(s_1 + \pi - x_2, \kappa_{12}) = \mathcal{M}\left(s_1 - \Delta\hat{s}_1, \Delta\hat{\kappa}_1\right), \tag{40}$$

where the mean and concentration of the posterior of the resdidual signals in the form of von Mises distribution are

$$\Delta\hat{s}_1 = \text{atan2}(\kappa_1 \sin x_1 - \kappa_{12} \sin x_2, \kappa_1 \cos x_1 - \kappa_{12} \cos x_2), \tag{41}$$

$$\Delta\hat{\kappa}_1 = \left[\kappa_1^2 + \kappa_{12}^2 - 2\kappa_1\kappa_{12}\cos(x_1 - x_2)\right]^{1/2}. \tag{42}$$

This is exactly parallel to that of the $\hat{s}_1$ and $\hat{\kappa}_1$. The implication is that in each module, the probability distributions of consensus and the residual informations might be explicitly represented in neural activities.

The above equations on the means and the variance of the posterior distributions of the general consensus and individual module's residual information provide the Bayes optimal estimate of the model, which we will compare against to what we can read out from the neuronal activities of the neural circuit model. This is a criterion for evaluating whether the neural circuit model is performing optimal Bayesian model selection as we wish.

4.1.5 Model Selection: integration versus segregation

Now we consider the two models (priors), integration and segregation, for explaining the observed inputs x_1 and x_2

$$p(s_1, s_2|M_{int}) = \frac{1}{(2\pi)^2 I_0(\kappa_s)} \exp\left[\kappa_s \cos(s_1 - s_2)\right], \tag{43}$$

$$p(s_1, s_2|M_{seg}) = p(s_1)p(s_2) = \frac{1}{(2\pi)^2} \tag{44}$$

The integration prior is the same as Eq. (30) in above. The segration prior can be seen as a special case when $\kappa_s = 0$, and then the two stimuli, s_1 and s_2, are independent of each other.

In the case of only two observation channels, the posterior of two models given two observations x_1 and x_2 can be calculated by using Bayes theorem,

$$\begin{aligned} p(M_{seg}|x_1, x_2) &= \frac{p(x_1, x_2|M_{seg})p(M_{seg})}{p(x_1, x_2)}, \\ &= \frac{p(x_1, x_2|M_{seg})p(M_{seg})}{p(x_1, x_2|M_{int})p(M_{int}) + p(x_1, x_2|M_{seg})p(M_{seg})}, \\ &= \frac{1}{1 + \frac{p(x_1,x_2|M_{int})}{p(x_1,x_2|M_{seg})} \frac{p(M_{int})}{p(M_{seg})}}, \end{aligned} \tag{45}$$

where

$$p(x_1, x_2|M) = \int p(x_1, x_2|\mathbf{s})p(\mathbf{s}|M)d\mathbf{s}. \tag{46}$$

In Bayesian statistics, $p(x_1, x_2|M)$ is called the *model evidence*, because it measures the explained probability of observations x_1 and x_2 by using a particular model M. $p(x_1, x_2|M_{int})/p(x_1, x_2|M_{int})$ is called the *Bayes factor*, which measures the likelihood ratior of two models on explaining cues.

In the network, we seek to develop a way to read out the Bayes factor directly from congruent and opponent neurons, which together maintain all the information of the observations x_1 and x_2.

In order to derive the equations for these read-out, let us first calculate the expression of $p(x_1, x_2|M)$. For integration model, we have

$$\begin{aligned} p(x_1, x_2|M_{int}) &= \int p(x_1, x_2|s_1, s_2)p(s_1, s_2|M_{int})ds_1 ds_2, \\ &= \frac{1}{(2\pi)^2 I_0(\kappa_x)} \exp\left[\kappa_x \cos(x_1 - x_2)\right], \end{aligned} \tag{47}$$

where $A(\kappa_x) = A(\kappa_1)A(\kappa_2)A(\kappa_s)$. $A(\kappa) = I_1(\kappa)/I_0(\kappa)$, while $I_1(\kappa)$ and $I_0(\kappa)$ are modified Bessel functions with first kind at first and zero order respectively. $A(\kappa)$ measures the *mean resultant length* of a distribution, which corresponds to the trigonometric mean of a circular variable. When κ_1, κ_2 and κ_s are sufficiently large, by using its Gaussian analogy, we can get

that $\kappa_x^{-1} \approx \kappa_1^{-1} + \kappa_2^{-1} + \kappa_s^{-1}$, and $\kappa_2^{-1} + \kappa_s^{-1}$ is equal to κ_2^{-1}, which is the total synaptic input from module 2 to module 1. This inverse of the concentration is analogous to the variance in a Gaussian concentration, indicating the reliability of the signals – the stronger the input from module 2, the more reliable (confident) it is, and the more neurons in module 1 and the model selection circuit should pay more weight to it. κ_x^{-1} is confidence of the network in explaining x_1, x_2 using the integration model ($p(x_1, x_2|M_{int})$) , it reflects the uncertainty about the input by choosing the particular prior model $p(s_1, s_2)$.

For the segregation model, we have

$$\begin{aligned} p(x_1, x_2|M_{seg}) &= \int p(x_1, x_2|s_1, s_2)p(s_1, s_2|M_{seg})ds_1 ds_2, \\ &= \int p(x_1|s_1)p(s_1)ds_1 \int p(x_2|s_2)p(s_2)ds_2, \\ &= \frac{1}{(2\pi)^2}. \end{aligned} \tag{48}$$

Here, $\kappa_x = 0$.

Finally, the Bayes factor for deciding whether to bind the two modules is given by,

$$\frac{p(x_1, x_2|M_{int})}{p(x_1, x_2|M_{seg})} = \frac{1}{I_0(\kappa_x)} \exp\left[\kappa_x \cos(s_1 - s_2)\right]. \tag{49}$$

Reading out κ_x is not trivial. Essentially, as we will later show, it requires the computation of product of κ_1, κ_2, the total synaptic input to the two modules, and the sum of $\kappa_1 + \kappa_{12}$. These signals can be computed by "normalization" neurons, because they sum up the all the synaptic inputs. The largest the responses are these neurons, the more confident are the signals they represent or normalizing.

4.2 Neural encoding model: probabilistic population code of visual attribute

To implement the abstract probabilistic model in the last section in neural circuit model, we need to represent the probability distribution dynamically in terms of stochastic neuronal population activities. Before going into the detailed neural circuit model, we will discuss how probability distributions are encoded in neuronal population activities.

The posterior distribution $p(\mathbf{s}|\mathbf{x})$ is high dimensional. For illustrative purpose, we firstly present how to represent each stimulus posterior $p(s_i|\mathbf{x})$ by a neural population, and then about how to represent the correlation between s_i and s_j between neural populations.

4.2.1 Representation of $p(s_i|\mathbf{x})$ in a neural population

We adopt a widely-used encoding model that $p(s_i|\mathbf{x})$ is represented by the firing activities of N neurons which are independent with each other, and each satisfies a Poisson distribution with the rate specified by its tuning curve [30]. Following is the encoding model for s_1 (the case for other s_i is similar),

$$\begin{aligned} \ln p(\mathbf{r}_1|s_1) &= \ln\left[\prod_{i=1}^{N} p\left(r_1(i)|s_1\right)\right], \\ &= \sum_{i=1}^{N} \ln\left[\frac{f_i(s_1)^{r_1(i)}}{r_1(i)!} e^{-f_i(s_1)}\right], \\ &= \sum_{i=1}^{N} r_1(i) f_i(s) - \sum_{i=1}^{N} f_i(s_1) - \sum_{i=1}^{N} \ln\left(r_1(i)!\right), \end{aligned} \tag{50}$$

where $r_1(i)$ and $f_i(s_1)$ are the firing rate and tuning curve of i-th neuron representing s_1, respectively. Because heading direction is a circular variable ranging from $-\pi$ to π, the tuning curve will be modeled as a circular function,

$$f_i(s_1) = R \exp\left[a \cos(\theta_i - s_1)\right], \tag{51}$$

where R is the maximal firing rate of the neuron, and θ_i is the preferred direction. Please note that the opponent value will be undefined if the tuning curve is modeled as a Gaussian function, because it is not periodic. With the assumption that the summed activities of whole neural population (second term in Eq. 50) is a constant and irrelevant to the stimulus value, and discarding the irrelevant terms, we can get the detailed expression for the encoding model

$$\ln p(\mathbf{r}_1|s_1) \;=\; a \sum_{i=1}^{N} r_1(i) \cos(\theta_i - s_1). \tag{52}$$

And then the distribution for stimulus s_1 becomes a von Mises distribution [31],

$$\begin{aligned} p(s_1|\mathbf{r}_1) &= \mathcal{M}[s_1; \hat{s}_1, \hat{\kappa}_1], \\ &= \frac{1}{2\pi I_0(\hat{\kappa}_1)} \exp[\hat{\kappa}_1 \cos(s_1 - \hat{s}_1)], \end{aligned} \tag{53}$$

where $I_0(\kappa) = (2\pi)^{-1} \int_{-\pi}^{\pi} e^{\kappa \cos(\theta)} d\theta$ is the modified Bessel function of the first kind and order zero. The mean $\hat{s}_1$ and concentration $\hat{\kappa}_1$ of stimulus are calculated to be,

$$\begin{aligned} \hat{s}_1 &= \text{atan2}\left(\textstyle\sum_{i=1}^{N} r_1(i) \sin\theta_i, \sum_{i=1}^{N} r_{s_1}(i) \cos\theta_i\right), \\ \hat{\kappa}_1 &= \left[\left(\textstyle\sum_{i=1}^{N} r_1(i) \sin\theta_i\right)^2 + \left(\sum_{i=1}^{N} r_1(i) \cos\theta_i\right)^2\right]^{1/2}, \end{aligned} \tag{54}$$

where atan2 is the arctangent function with two arguments, which calculates the angle of a vector. For the von Mises distribution, the concentration parameter $\hat{\kappa}_1$ is usually a positive number and measures the dispersion of distribution. A larger $\hat{\kappa}_1$ means a more concentrated distribution, and hence a more reliable the estimate is. In the large κ limit, by expanding the cos function in von Mises distribution up to second order, i.e., $\cos(x - s) \approx 1 - (x - s)^2/2$, a von Mises distribution $\mathcal{M}[\mu, \kappa]$ approaches to a normal distribution $\mathcal{N}[\mu, \kappa^{-1}]$ with variance denoted by κ^{-1}.

4.2.2 Representation of posterior $p(s_i|\mathbf{x})$

Given the encoding model, we then investigate what kind of operations are needed for the neural representation to implement information integration. Because the stimulus s_1 is fully represented by the neural population $\mathbf{r_1}$, the activities of the neural population should satisfy

$$\ln p(\mathbf{r}_1|x_1, x_2) = \ln p(\mathbf{r}_1|x_1) + \ln p(\mathbf{r}_1|x_2), \tag{55}$$

which is the same as the abstract model, except that s_1 has been replaced with r_1. Substituting the encoding model (Eq. 52) into above equation, we can find that

$$r_{1|x_1,x_2}(i) = r_{1|x_1}(i) + r_{1|x_2}(i), \tag{56}$$

where $r_{1|x_1,x_2}(i)$ denotes the firing rate of i-th neuron representing s_1 given the cues x_1 and x_2 together. The above equation tells us that the firing rate given two cues should be a linear sum of the firing rate when given either cues, which is the same as in previous work [29].

4.2.3 Representing the residual information

We can also substitute the encoding model (Eq. 52) into definition of residual information (Eq. 39) to see what kind of operations are needed:

$$\sum_i r_{1|x_1,x_2}(i)\cos(\theta_i - s_1) = \sum_i r_{1|x_1}(i)\cos(\theta_i - s_1) - \sum_i r_{1|x_2}(i)\cos(\theta_i - s_1). \tag{57}$$

At first sight, above equation indicates that the multisensory divergence can be achieved by the suppression from the neural activity when giving cue 2, i.e., $r_{1|x_1,x_2}(i) = r_{1|x_1}(i) - r_{1|x_2}(i)$. However, due to the constraint that the neuronal firing rate is a positive number, the $r_{1|x_1,x_2}(i)$ would be rectified to be zero if $r_{1|x_2}(i)$ is larger than $r_{1|x_1}(i)$. When this happens, the neurons cannot express how much the residual between two inputs, because the firing rates are all zero, and there is a large distortion due to this neuronal constraint.

This problem can be resolved by using one property of cosine function that $\cos(x+\pi) = -\cos(x)$. And then the suppression by the neural activity given cue 2 (the last term in Eq. 57) can be transformed into a facilitation from the same activity but targeted on the neurons in the opponent direction, i.e.,

$$-\sum_i r_{1|x_2}(i)\cos(\theta_i - s_1) = \sum_i r_{1|x_2}(i)\cos[(\theta_i + \pi) - s_1)]. \tag{58}$$

This transformation has clear consequences with regards to the neural tuning curves. For the ith neuron in the population, we can see its preferred direction for cue 1 is θ_i; while the preferred direction under cue 2 shifts to opponent direction $\theta_i + \pi$. Thus, this neuron is an opponent neuron, according to the definition.

4.2.4 Reading out the Bayes factor

Now we investigate how we could read out the Bayesa factor (Eq. 49) in neural population activities. Reorganizing the terms in Eq. (49), we have

$$\begin{aligned}\frac{p(x_1,x_2|M_{int})}{p(x_1,x_2|M_{seg})} &= \exp\left[\kappa_x\cos(x_1-x_2) - \ln I_0(\kappa_x)\right], \\ &\approx \exp\left[\kappa_x\cos(x_1-x_2) - \left(\kappa_x - \frac{1}{2}\ln(2\pi\kappa_x)\right)\right]\end{aligned} \tag{59}$$

The approximation is from the fact that $I_0(\kappa) \approx \kappa/\sqrt{2\pi\kappa}$, which can be obtained from the analogy between the von Mises distribution and the Gaussian distribution.

Reading out κ_x

In order to derive a neural implementation of above equation, let us take a look again the physical meaning of the terms. From the calculations above, κ_x is derived to be

$$\begin{aligned}\kappa_x^{-1} &\approx \kappa_1^{-1} + \kappa_2^{-1} + \kappa_s^{-1}, \\ &\approx \kappa_1^{-1} + \kappa_{12}^{-1},\end{aligned} \tag{60}$$

where $\kappa_{12}^{-1} = \kappa_2^{-1} + \kappa_s^{-1}$, and κ_{12} is the concentration of $p(s_1|x_2)$. Considering the analogy between the von Mises distribution and the Gaussian distribution, κ^{-1} corresponds to the variance of Gaussian distribution. Taking the reciprocal of above equation, we have

$$\kappa_x = \frac{\kappa_1\kappa_{12}}{\kappa_1 + \kappa_{12}} \tag{61}$$

We see that κ_x has a divisive normalization form with respect to inputs κ_1 and κ_{12}. κ_1 and κ_{12} have clear meanings in neural population activities (Eq. 54),

- κ_1 is the concentration of likelihood $p(x_1|s_1)$, and thus is represented by the summed activities of feedforward inputs received by congruent neurons in network module 1.

- κ_{12} is the concentration of indirect likelihood $p(x_2|s_1)$, and can be mapped with the summed activities of horizontal inputs emitted from congruent neurons in network module 2 and received by congruent neurons in network module 1.

Considering divisive normalization is widely observed in cortical circuits, it is possible that κ_x can be read out by taking a divisive normalization of feedforward inputs and the reciprocal inputs from another network module. Once κ_x can be read out, the terms left inside exponential function in Eq. (54) can be considered as a function of κ_x. Next, we consider how to read them out from neuronal population activities.

Reading out $\kappa_x \cos(x_1 - x_2)$

Substituting the divisive normalization form of κ_x (Eq. 61) into Eq. (54), the first term in Eq. (59) becomes

$$\kappa_x \cos(x_1 - x_2) = \frac{\kappa_1 \kappa_{12} \cos(x_1 - x_2)}{\kappa_1 + \kappa_{12}}. \tag{62}$$

How do we find $\cos(x_1 - x_2)$, the residuals between two input streams, from the neural population response? Recall that the concentration of the posterior, $\hat{\kappa}_1$ (Eq. 37), and the concentration of residual information, $\Delta\hat{\kappa}_1$ (Eq. 42), are both functions of $\cos(x_1 - x_2)$. And thus $\cos(x_1 - x_2)$ can be read out by their difference given by

$$\kappa_1 \kappa_{12} \cos(x_1 - x_2) = \frac{1}{4}\left(\hat{\kappa}_1^2 - \Delta\hat{\kappa}_1^2\right), \tag{63}$$

$\hat{\kappa}_1$ and $\Delta\hat{\kappa}_1$ have clear meanings in neural population activities.

- $\hat{\kappa}_1$ is the concentration of the posterior $p(s_1|x_1, x_2)$, and is represented by the summed activities of congruent neurons (Eq. 54).

- $\Delta\hat{\kappa}_1$ is the concentration of residual information $D(s_1|x_1; s_1|x_2)$, and is represented by the summed activities of opponent neurons.

The above equation suggests that the difference between the activities of congruent and opponent neurons encode the residual signal between the two modules' consensus and their individual isolated decisions.

Finally, we obtain that

$$\kappa_x \cos(x_1 - x_2) = \frac{\hat{\kappa}_1^2 - \Delta\hat{\kappa}_1^2}{4(\kappa_1 + \kappa_{12})}. \tag{64}$$

It indicates $\kappa_x \cos(x_1 - x_2)$ can be read out from the difference of activities between congruent and neurons and divisively normalized by the summation of feedforward and horizontal inputs.

Reading out $\kappa_x - \ln(2\pi\kappa_x)/2$

The second term in Eq. (59) is a function with respect to κ_x

$$\kappa_x - \frac{1}{2}\ln\left(2\pi\kappa_x\right). \tag{65}$$

When plotting this term with respect to κ_x, the curve looks like the lower part of a sigmoid funcion. Thus, this function may be implemented as the input-output nonlinearity of a type of neurons in cortex.

The analyses in this section aims to give us some insight on how to design a network model to read out the Bayes factor and then use it to achieve integration and segregation. Next, we explore a neural network model to achieve the desired functionality of integration and segregation.

4.3 Neural network model

We also propose a network model to implement this Bayesian model selection. In order to process two cues, the neural network model consists of two interconnected network modules which each receives the input from a cue. For illustrative purposes, the network model is decomposed into three functional parts to help readers understand the function of each part:

1. Posterior
2. Residual information
3. Model selection circuit

The network model consists of two parts (Fig. 2A): 1) two interconnected network form a distributed representation of posterior and residual information (Eq. 21); and 2) a model selection circuit between two networks to estimate the probability of integration/segregation (Eq. 45), whose activities gate the horizontal connections between two networks and thus determines whether to integrate two inputs together.

Each network can be regarded as a module in a brain area, and hence, the two interconnected networks can regard as two modules within the same brain areas or across different areas. The model selection circuit includes some interneurons which gate the excitatory horizontal connections. Below, we will briefly introduce the structure of each part.

4.3.1 Distributed representation network

Each network has a group of congruent (C) neurons and a group of opponent (O) neurons with the same number of neurons. Both groups of neurons receive feedforward inputs from cues, and crosstalk with their counterparts in another network through horizontal connection (green and yellow arrows in Fig. 2A). Each group of neurons has an inhibition pool to provide divisive normalization over neuronal activities (filled red circle in Fig. 2A).

Within a group of neurons, each neuron has a preferred stimulus and bell-shaped tuning curve. The neurons within the same group are interconnected together and the connection strength between two neurons is a bell-shape function with regards to their tuning similarity, i.e. two neurons will be more tightly connected if their tunings are more similar. This kind of recurrent connection enables the neurons to form a continuous attractor neural network (CANN) [14, 32] (Fig. 2B), which is widely used to explain the coding of continuous stimuli in our brain, such as orientation, moving direction, or spatial location. It has been proved that a CANN can approximately achieve maximal likelihood estimate of its input [33, 34].

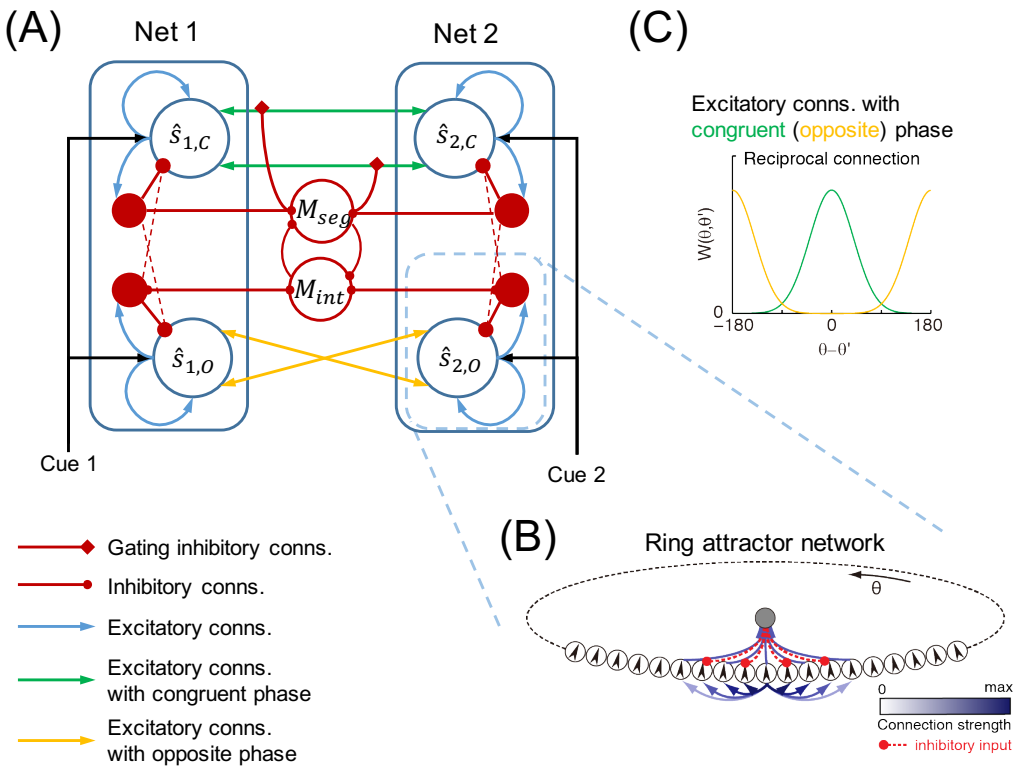

Figure 2: Network architecture of Bayesian model selection.

The congruent and opponent neurons are named by their different tuning properties with respect to two cues, which results from different lateral connections between networks in our model. Across two networks, congruent neurons are connected with the same orientation (green line in Fig. 2A&C), meaning two neurons with similar preferred stimulus will be tightly connected. And then the neurons have similar preferred stimulus value given two cues (Fig. 2A). In contrast, opponent neurons of orthogonal orientations are connected across modules (hypercolumns) (yellow line in Fig. 2C). For example, an opponent neuron in network 1 which prefers 0° of cue 1 will most tightly connect with an opponent neurons in network 2 and prefers 90° orientation of cue 2. And thus the preferred stimulus of an opponent neurons under two cues are opponent (Fig. 2B).

4.3.2 Model selection circuit

The model selection circuit consists of two competing inhibitory neuron pools (open red circles in Fig. 2A). One neuronal pool represents the selection of the integration model M_{int} and the other represents the selection of the segregation model M_{seg}. The M_{seg} neuronal pool gates the horizontal connections between two networks, which might be achieved through shunting inhibition in cortical circuit [35]. The M_{int} neurons receives inhibitory inputs from inhibition pools (summed activities of all neurons) associated with opponent neurons from two networks. In contrast, the M_{seg} receives inhibitory inputs from inhibition pools of congruent neurons.

Intuitively, when two cues are similar, the firing rate of congruent neurons will be larger than

that of the opponent neurons. Then, the M_{seg} neuronal pool will receives stronger inhibition than M_{int}, and thus M_{int} will have larger probability to win the competition. In this case, the response of M_{seg} is completely inhibited by M_{int} and then two networks will communicate with each other, which is equivalent to selecting the integration model. In contrast, when two cues are different, M_{seg} receives smaller inhibition and has larger probability to win. Then, the activity of M_{seg} will cut the horizontal connections between two networks, corresponding to selecting segregation prior.

We introduced this mechanism, using the recurrent competitive decision circuit of Xiao-Jing Wang [36] as an alternative mechanism for grouping and segmentation. Rather than inferring a scaling variable to indicate or bind the feature detectors to a selected group as in Schwartz's model, we compute residuals and general consensus of the modules, and use this information to decide whether a particular module belongs to the group of global consensus.

4.4 Mathematical Description of Neural Circuits

Now, we present the detailed dynamical equations for the neural circuit model, which is a type of dynamical-system based neural network models with firing-rate based neurons.

4.4.1 The model dynamics

$u_{m,n}(\theta)$ and $r_{m,n}(\theta)$ denoted the synaptic input and firing rate, respectively, of a n-type neuron in module m, whose preferred orientation with respect to the direct cue m is θ. $n \in \{c, o\}$ represents the congruent and opponent cells, respectively, and $m \in \{1, 2\}$ is the index of a network. Each network module can be regarded to a module in V1. For simplicity, we assume that the two network modules are symmetric and only present the dynamics of module 1. The dynamics of network module 2 can be obtained by simply replacing the indices 1 and 2.

4.4.2 Congruent neurons

The dynamics of a congruent neuron in network module (module) 1 is given by

$$\begin{aligned} \tau \frac{\partial u_{1,c}(\theta,t)}{\partial t} &= -u_{1,c}(\theta,t) + I_{1,c}(\theta,t) \\ &\sum_{\theta'=-\pi}^{\pi} W_r(\theta,\theta') r_{1,c}(\theta',t) + g(x_1,x_2) \sum_{\theta'=-\pi}^{\pi} W_c(\theta,\theta') r_{2,c}(\theta',t) \end{aligned} \tag{66}$$

where τ is the time constant for the synaptic input. For simplicity, we set $\tau = 1$. $I_{1,c}(\theta,t)$ is the feedforward input to the neuron from the lateral geniculate nucleus (LGN), conveying contrast information. $W_r(\theta,\theta')$ is the recurrent connections from congruent neuron prefering orientation θ' to another congruent neuron θ within the same network module. Finally, $W_c(\theta,\theta')$ is the weight linking two congruent neurons across network modules, from congruent θ' in network 2 to congruent neuron θ in network 1. We assume that there are no excitatory connections between congruent and opponent neurons. $g(x_1,x_2)$ is a gating variable, which gates the connections across two network modules. It is a function with respect to the orientation x_1 and x_2 received by two network modules.

Feedforward input

The feedforward input conveys the direct cue information to a module (e.g. the feedforward input fed into a module in V1 which has the information of orientation), which is

$$I_{1,n}(\theta,t)=\alpha_1\exp\Big[\frac{\kappa_a}{2}\cos(\theta-x_1)\Big]+\sqrt{F_1\exp\Big[\frac{\kappa_a}{2}\cos(\theta-x_1)\Big]}\xi_1(\theta,t)+I_b+\sqrt{F_1 I_b}\epsilon_{1,n}(\theta,t), \tag{67}$$

where α_1 is the input intensity, I_b is the mean of background input, and F is the Fano factor. κ_a is the tuning width of input. x_1 represents the orientation of visual stimulus, and controls the position of the bump input $I_{1,n}$ in feature space. $\xi_1(\theta,t)$ and $\epsilon_{1,n}(\theta,t)$ are Gaussian white noise with zero mean and variance satisfying $\langle\xi_m(\theta,t)\xi_{m'}(\theta',t')\rangle=\delta_{mm'}\delta(\theta-\theta')\delta(t-t')$, $\langle\epsilon_{m,n}(\theta,t)\epsilon_{m',n'}(\theta',t')\rangle=\delta_{mm'}\delta_{nn'}\delta(\theta-\theta')\delta(t-t')$. The signal-associated noises $\xi_1(\theta,t)$ to congruent and opponent neurons are exactly the same, while the background noises $\epsilon_{1,n}(\theta,t)$ to congruent and opponent neurons are independent of each other. At the steady state, the signal drives the network state to center at the cue value x_1, whereas noises induce fluctuations of the network state. Since we consider multiplicative noise with a constant Fano factor, the signal strength α_m controls the reliability of cue m [29], which is derived from a neural encoding model (Eq. 54). The exact form of the feedforward input is not crucial, as long as it has a uni-modal shape.

Connection matrix

$W_r(\theta,\theta')$ is the recurrent connection between neurons in the same module, which is a von Mises function with respect to the difference in tunings between two neurons,

$$W_r(\theta,\theta')=\frac{J_r}{2\pi I_0(\kappa_a)}\exp\big[\kappa_a\cos(\theta-\theta')\big] \tag{68}$$

where κ_a controls the tuning width of the congruent neurons. $I_0(\kappa_a)$ is a modified Bessel function of the first kind, to ensure the integral of $\int W_r(\theta,\theta')d\theta'=J_r$. $W_c(\theta,\theta')$ is the reciprocal connection between congruent cells in two modules, which is set to be

$$W_c(\theta,\theta')=\frac{J_c}{2\pi I_0(\kappa_a)}\exp\big[\kappa_a\cos(\theta-\theta')\big]. \tag{69}$$

The reciprocal connection strength J_c controls the extent to which cues are integrated between modules and is associated with the correlation parameter κ_s in the stimulus prior (see abstract probability distribution in the above).

4.4.3 Opponent neurons

Similar with congruent neurons, the dynamics of an opponent neuron in module 1 is given by

$$\tau\frac{\partial u_{1,o}(\theta,t)}{\partial t}=-u_{1,o}(\theta,t)+\sum_{\theta'=-\pi}^{\pi}W_r(\theta,\theta')r_{1,o}(\theta',t)+\sum_{\theta'=-\pi}^{\pi}W_o(\theta,\theta')r_{2,o}(\theta',t)+I_{1,o}(\theta,t). \tag{70}$$

It has the same form as that of a congruent neuron except that the pattern of reciprocal connections linking opponent neurons across two network modules are given by

$$W_o(\theta,\theta')\;=\;\frac{J_c}{\sqrt{2\pi}a}\exp\big[\kappa_a\cos(\theta+\pi/2-\theta')\big]=W_c(\theta+\pi/2,\theta'). \tag{71}$$

That is, opponent neurons between modules are opponently connected by an offset of $\pi/2$, which is the half period of orientation π. We choose the strength and width of the connection pattern W_o to be the same as that of W_c. This is based on the finding that the tuning functions of congruent and opponent neurons have similar tuning width and strength [37].

4.4.4 Divisive normalization

In the model, we include the effect of inhibitory neurons through a divisive normalization (DN) to the responses of excitatory neurons [38]. In a network module, congruent and opponent neurons have their own DN neurons, which receive the inputs from corresponding excitatory neurons from two network modules. For example, the firing rate of neurons with type $n = c, o$ in network 1 is given by

$$r_{1,n}(\theta, t) = \frac{[u_{1,n}(\theta, t)]_+^2}{1 + \omega D_{1,n}(t)}, \tag{72}$$

$$D_{1,n}(t) = \sum_{\theta'=-\pi}^{\pi} \left[u_{1,n}(\theta', t)\right]_+^2 + w_c^{(D)} \sum_{\theta'=-\pi}^{\pi} \left[u_{2,n}(\theta', t)\right]_+^2 \tag{73}$$

where ω controls the magnitude of divisive normalization, and $[x]_+ \equiv \max(x, 0)$ is a negative rectify function. $D_{1,n}(t)$ regards to the response of the inhibitory neuron in network 1 which sums up all activities of type n neurons in both networks. $w_c^{(D)}$ controls the relative weight of receiving summed activities of excitatory neurons across networks (the weight of receiving summed excitatory neuronal activities within the same network is 1).

4.5 Model selection circuit

The model selection circuit consists of two inhibitory neuronal pools, which inhibit with each other and thus exhibit a winner-take-all behavior. The winning of either neuronal pool represents the choice of either an integration or a segregation model (Eqs. 43 and 44). Denote u_{int} (r_{int}) and u_{seg} (r_{seg}) the synaptic inputs (firing rate) of integration and segregation neurons, respectively. Their dynamics are governed by

$$\tau \frac{d}{dt} u_{int}(t) = -u_{int}(t) - J_I r_{seg}(t) + I_{int}(t), \tag{74}$$

$$\tau \frac{d}{dt} u_{seg}(t) = -u_{seg}(t) - J_I r_{int}(t) + I_{seg}(t), \tag{75}$$

where τ is the time constant for the inhibitory neurons, J_I is the inhibitory connection strengh between two pools. For simplicity, we assume that the inhibitory connection strength across two pools is the same.

The input to firing rate curve of two inhibitory neuronal pools is

$$r_i(t) = \frac{[u_i(t)]_+^2}{1 + \omega_I [u_i(t)]_+^2}, \quad i = int,\ seg \tag{76}$$

The detailed shape of the input to firing rate curve is not critical, as long as it resembles the lower part of a sigmoid fucntion.

I_{int} and I_{seg} are the inputs fed into integration and segregation inhibitory neurons, respectively. The integration neuron received inhibitory inputs from divisive normaliztion neurons

associated with opponent neurons in two network modules; while the segregation neurons receives inhibitory inputs from normalization neurons of congruent neurons in two networks. Mathematically, I_{int} and I_{seg} satisfy

$$I_{int}(t) = D_{1,o}(t) + D_{2,o}(t) + \sigma_{int}\xi_{int}(t), \tag{77}$$

$$I_{seg}(t) = D_{1,c}(t) + D_{2,c}(t) + \sigma_{seg}\xi_{seg}(t), \tag{78}$$

where $\xi_{int}(t)$ and $\xi_{seg}(t)$ are the background noise corrupting the input, and their noise strength is specified by σ_{int} and σ_{seg}.

Previous studies demonstrate the winning probability of one neuronal pool, e.g., M_{int}, is approximately a sigmoid function with respect to the input difference received by two neuronal pools,

$$p(M_{int}|x_1, x_2) = \frac{1}{1 + \exp(-\beta(I_{int} - I_{seg})}. \tag{79}$$

β measures the noisiness of the model selection circuit, which is a function of input intensity, input noise strength, and mutual inhibition strength in model selection circuit.

The winning probability of one neural pool in the model selection circuit is probabilistic, which matches the probability predicted by Bayes theorem, called probability matching [39]. In a strict Bayesian model selection approach, if $p(M_{int}|\mathbf{x}) = 0.7$ ($p(M_{seg}|\mathbf{x}) = 0.3$), then M_{int} is always being selected. In contrast, in probability matching, it means that the M_{int} has 70% chance of being selected, corresponding M_{int} neuronal pool will win 70% of all trials. Psychophysical experiments have found that probability matching is adopted by most subjects [39].

When r_{seg} wins, it will break the connections across two network modules, i.e.,

$$g(x_1, x_2) = \begin{cases} 0, & \text{if } r_{seg} \text{ wins,} \\ 1, & \text{if } r_{int} \text{ wins.} \end{cases} \tag{80}$$

5 Experimental Results

First, we will describe the basic working of the model to evaluate model fitness – the extent to which the network, as a dynamical system, operates in the way that we propose it will. The criteria include the reading out of the Bayes factor, the working of the model selection decision process, and the correct computation of the optimal Bayes estimates of the global consensus (congruent neurons) and the residuals (opponent neurons). We will provide a demonstration of the operation of each component of the model.

Next, we will use a set of neural findings to test the model and to compare it with other current models. The primary neural findings we seek to explain are our novel findings on the development of familiarity effect of global object image patterns in early visual cortex due to statistical learning as documented in Section 8.1,8.2.

We will relate the different components of the model to well-known circuit motifs and cell types to provide insights into the potential computations of real neural circuits.

5.1 Fitness of the model in Bayesian model selection

5.1.1 Representations of input, consensus, and residual distributions

The architecture of the implementation of a two-module network is shown in Fig. 2. The inputs to the network are provided by $\vec{x}_1$ and $\vec{x}_2$. In our example, each element of these input vectors will be the output of an oriented filter (e.g rectified Gabor filter) in response to an image pattern. Each element of $\vec{x}_1$ is fed into a neuron in the congruent $\vec{\hat{s}}_{1,c}$ pool and a neuron in the opponent $\vec{\hat{s}}_{1,o}$ pool in module 1. Similarly, each element of $\vec{x}_2$ is input into the neuron in the $\vec{\hat{s}}_{2,c}$ and $\vec{\hat{s}}_{2,o}$ pools in module 2. We use **90** different orientation selective neurons in each pool to span the orientation parameter space smoothly. The input is noisy and its average over time is a Gaussian bump centered at the orientation of the input visual attribute as shown in the first row of Fig. 3.

Congruent neurons of similar orientation tunings are connected together across the modules with excitatory connections, consistent with known iso-orientation facilitatory connections. Opponent neurons are connected with opponent neurons of orthogonal orientation preferences across the two modules. This model predicts the existence of opponent neurons which show surround suppression and could be excited by stimuli of orthogonal orientation in the surround.

Fig. 3 shows the input activity x_1 over 50 time units, as well as the population activity of the congruent and opponent neurons (left panel). The middle column shows the mean activity of x_1, $s_{1,c}$, $s_{1,o}$, respectively. The variance of $s_{1,c}$, $s_{1,o}$ are shown, while the variance of the x_1 is insignificant and is not shown. In this case, x_1, and x_2 indicate the same visual attribute – the mean of the Gaussian bumps centered on the same orientation, and hence the two modules are in complete agreement. The neural dynamics of the recurrent network only serve to clean up and stabilize the inference solution.

The activity $\hat{\kappa}_{1,c}$ of the congruent normalization neurons is mostly determined by the recurrent activation of the active congruent neurons. This is the input going into the segregation neurons of the decision circuit. As shown in the last row of Fig. 3, because s_1 and s_2 are in agreement, overall the congruent neurons will respond more strongly than the opponent neurons, resulting in greater activation of the congruent normalization neurons relative to the opponent normalization neurons. This will drive the decision circuit to favor the integration model. On the other hand, when x_1 is significantly different from x_2, the opponent normalization neurons will have a stronger response than the congruent normalization neurons.

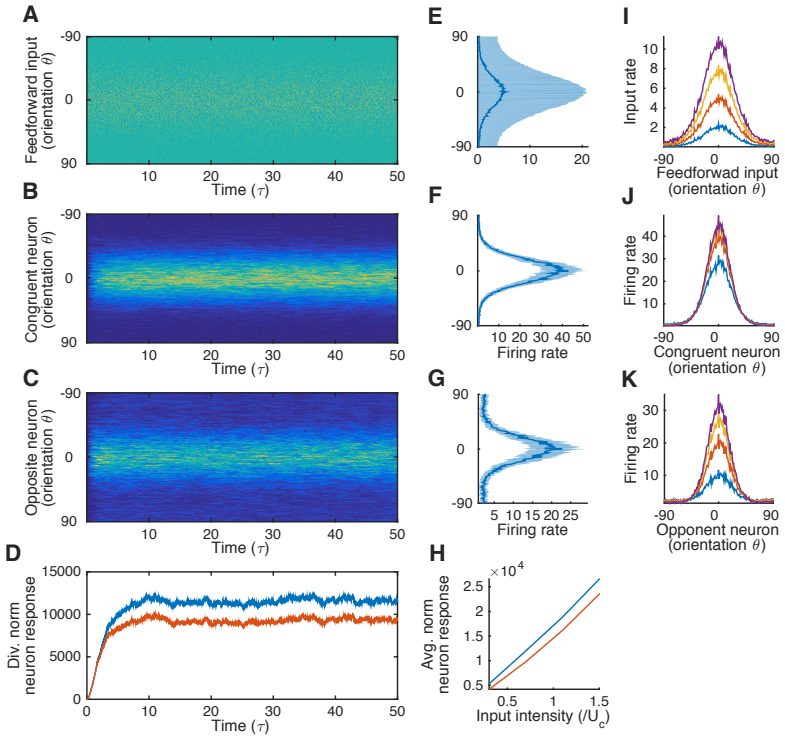

Figure 3: Left panel: Activity of the input (A), congruent (B) and opponent (C) neurons of module 1 over time are shown in the first three rows. The two modules coupled together under the integration mode is shown. D: Blue curve represents the activity of congruent normalization neuron and red curve is the activity of the opponent normalization neuron. The mean inputs to the two modules x_1, and x_2 are centered on the same orientation. Because of the agreement of the two congruent populations, the congruent neurons fire more strongly, resulting in greater responses of their congruent normalization neurons, over that of the opponent normalization neurons. This drives the decision circuit to favor integration. The normalization congruent neurons' activity reflects the increased confidence on the global consensus due to the coupling and agreement between the two modules. Middle column: (E, F, G) time average activity. Note the feedforward input's variance is very large (noisy) (E), which can reduced in (F) and (G). Right panel: The input of different orientations at four different contrasts (I), and tuning curve of the congruent neuron (J) and opponent neurons (K) at different contrasts.

The posterior of the global consensus is represented as $\vec{s}_{1,c}$ by the activities of the excitatory congruent neurons in module 1, and similarly $\vec{s}_{2,c}$ is represented by the congruent neurons in module 2. Without horizontal interaction, the global consensus represented by each module is just the "visual attribute" inferred by the individual module, cleaned up and stabilized by the recurrent attractor dynamics of the network. With horizontal connections, the global consensus is represented by the joint posterior $(\vec{s}_{1,c}, \vec{s}_{2,c})$, but $\vec{s}_{1,c}$ in each module is computed with the considerations of the messages from the other modules. Each neuron in $\vec{s}_{2,c}$ is connected to the neuron in $\vec{s}_{1,c}$ of the same or similar orientation, with Gaussian connection weights. Thus, congruent neurons with the same tunings cooperate and reinforce each other to arrive at a global consensus, which can be viewed as a "decentralized" representation of a global (semantic) concept.

The opponent neurons in each module help keep track of the residual signals of the module, which is the difference between the module's original opinions and the global consensus arrived through network interaction. These residual signals serve two purposes. First, they can be used to drive the segmentation process by inhibiting the integration neurons, and second, they can be used to synthesize or reconstruct the original input by simply combining the activities of the congruent and opponent neurons. There are multiple ways to implement the network computation of these residuals. Currently, we adopted a scheme that is known to work in the scenario of the cue integration [40] which allows excitatory interactions between neurons coding for the orthogonal orientations across different modules. This scheme works when we assume the neurons are coding classical visual attributes, such as orientation, motion direction, stereo disparity, and color. For more abstract, higher-order, or complicated concepts, some creative new schemes may need to be developed.

5.1.2 Computation of the Bayes factor and the model selection decision circuit

The decision circuit is shown in Fig. 2. The segregation neurons in the decision circuit receive inhibitory input $\hat{\kappa}_1^2$ which are the responses of the congruent normalization neurons while the integration neurons receive inhibitory input $\Delta\hat{\kappa}_1^2$ from the opponent normalization neurons. Thus, the congruent normalization neurons and the opponent normalization neurons of the two modules drive the segregation neurons and the integration neurons in the decision circuit respectively, forming an inhibiton-disinhibiton circuit. Since SOM neurons do not inhibit each other, the integration and segregation neurons are likely PV neurons. The segregation neurons, when active, should shut off the message passing between module 1 and module 2 neurons. If these PV segregation neurons are basket cells, they could form elaborate inhibitory nest around the target cell, and shut off the incoming signals from other modules by shunting the appropriate dendritic branches of the congruent neurons without turning off its responses. The integration neurons compete with the segregation neurons. When they win, they will also feed back to inhibit the opponent normalization neurons. This feedback to opponent normalization neurons is for synchronizing the normalization neurons across modules to increase the correlated activities of the opponent and congruent neurons. This feedback is mostly relevant for spiking model and is not implemented in our current model.

The horizontal or recurrent connections between the modules not only projects to each of their excitatory congruent neurons, but also converges on a surround-sum excitatory neuron to encode the total synaptic input from the module. This surround-sum neuron then projects to a SOM neuron whose response increases monotonically with an increase in iso-orientation surround. The SOM cells normalize the excitatory congruent or opponent neurons by contacting their distant dendrites to provide "surround modulation". They, together with the congruent normalization neurons, provide suppressive input to the segregation PV neuron – strong consensus and congruent activities will suppress the segregation and favors integration. However, when the segregation neuron wins, it deactivates the surround-sum neuron, thereby depriving the SOM neurons of their excitatory excitatory input from the surrounding modules. The decrease in "contextual" divisive normalization of the congruent neurons by SOM in an isolated module will lead to an increase in firing rates of the congruent neurons in that module.

The winning probability of the segregation neurons due to network dynamics from the previous sections is given by:

$$p(M_{seg}|x_1, x_2) = \frac{1}{1 + \frac{p(x_1,x_2|M_{int})}{p(x_1,x_2|M_{seg})} \frac{p(M_{int})}{p(M_{seg})}} \tag{81}$$

The Bayes factor

$$\begin{aligned}\frac{p(x_1,x_2|M_{int})}{p(x_1,x_2|M_{seg})} &= \exp\left[\kappa_x\cos(x_1-x_2)-\ln I_0(\kappa_x)\right], \\ &\approx \exp\left[\kappa_x\cos(x_1-x_2)-\left(\kappa_x-\frac{1}{2}\ln\left(2\pi\kappa_x\right)\right)\right] \end{aligned} \tag{82}$$

can be read out from the congruent and opponent neurons population activities, where

$$\kappa_x = \frac{\kappa_1\kappa_{12}}{\kappa_1+\kappa_{12}} \tag{83}$$

The numerator of the first term of the Bayes factor can be read out from,

$$\kappa_1\kappa_{12}\cos(x_1-x_2) = \frac{1}{4}\left(\hat{\kappa}_1^2-\Delta\hat{\kappa}_1^2\right), \tag{84}$$

where $\hat{\kappa}_1$ is the concentration of the posterior of the consensus and the $\Delta\hat{\kappa}_1$ is the concentration of the posterior of the residual. The concentrations which are the inverse of variance, are directly related to the maximum firing rates in the congruent population and in the opponent populations. The stronger the concentrations or firing rates, the stronger the reliability of the signals.

The second term $\kappa_x - \ln(2\pi\kappa_x)/2$ is the normalization factor of the probability model and serves as a threshold. Empirically, it can be approximated by the lower part of the sigmoid function with respect to κ_x. The computation of κ_x here is not easy. It requires the product of two quantities, k_1, the sum of feedforward synaptic input and k_{12}, the surround-sum neuron encoding the total horizontal synaptic input. One possible scheme for implementing this product is through dendritic computation where k_1 and k_{12} can be combined super-linearly, and then normalized by their sum in order to adapt the threshold. Given Jeff Lichtman's findings on the complex and elaborate axonal and dendritic connection patterns even for a single neuron, computing product of two activities (k_1 and k_{12}) using dendritic computation is plausible and is investigated in our sparse coding studies (see Other Works section).

When the input into the two modules are substantially different in orientation, for example by 45 degrees ($x_1 = 0, x_2 = 45$ degrees), the disparity between the inputs will cause the two congruent neurons to negotiate to reach a global compromised consensus. Fig. 4A shows the population activity of the congruent neurons of of module 1 in a single trial, with the red line indicating the mean posterior decoded orientation. Fig. 4C shows the population activity of the congruent neurons of module 1 (red curve) when they are in the segregation condition, and when they are in the integration condition, where the compromise with module 2 shifts the population activity to the right (blue curve). Fig. 4D shows the population activity of the opponent neurons exhibiting the opposite effect, with the residual's posteriors keeping track of the compromise they have made. The overall population activity of both population of neurons are higher for segregation condition than for the integration condition, as one can also infer from Fig. 4C . Fig. 4B shows the area under the population curve of all the congruent neurons (C) and that of all the opponent neurons (O) under the integration and segregation conditions. The responses for congruent neurons and opponent neurons both decrease under integration because each of their divisive normalization neurons now receive additional input from the other modules, κ_{12} for module 1 from module 2, and κ_{21} from module 2 to module 1.

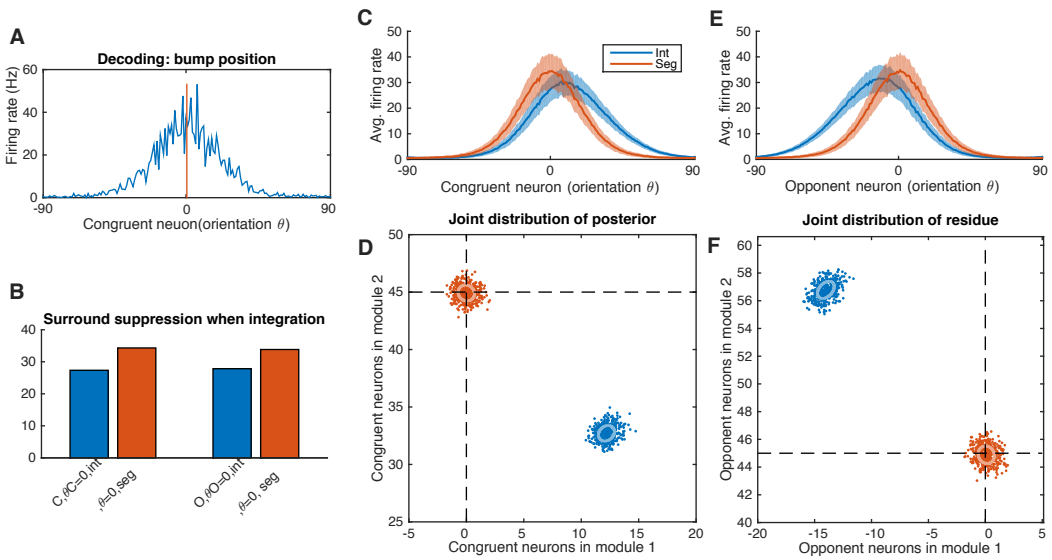

Figure 4: Representations of the posteriors of the global consensus and of the residues of module 1, with interaction between the two modules. Difference between the two input orientation is 45 degrees. A: the population activity of the congruent neurons in module 1, the red line denotes the position of population activity in feature space, which is interpreted as network's estimate (posterior). B: the firing rate of congruent and opponent neurons preferring 0° orientation in the center when integration (blue) and segregation (red) model is applied. C and E: mean firing rates of the congruent (C) and opponent E population in network module 1 when segregation model is chosen (red) and when integration model is selected (blue). D. The joint distribution of the estimate of congruent neurons in two network modules under integration (blue) and segregation (red) conditions. Dashed lines denote the input received by two networks. In the case of segregation, the two network modules don't cross-talk with each other, and thus the estimates of two networks' estimate are as the same as input (centered at the cross of two dashed lines). When integration happens, two networks communicate with each other, and then the distribution shifts to a direction where the two networks' estimates become more similar. F: the joint distribution from the opponent neurons in the two modules under integration (blue) and segregation (red). When integration happens, the joint distribution shifts to a direction which is opponent with the one of congruent neurons, which represents the residual information of input lost in congruent neurons.

Note that the overall congruent neurons' population activity is not always the same as that of the opponent neurons as shown in Fig. 4. When the input disparity between the two modules is small, the congruent neurons' activities will be higher than that of the opponent neurons, reflecting a higher confidence on the global consensus. When the input disparity between the two modules is large, the opponent neurons activity will be higher, indicating higher level of discord. The activities of the congruent normalization neurons will inhibit the segregation neurons in the decision circuit. Strong congruent activity will undermine the segregation neurons, and thus indirectly help the integration neurons by disinhibition. The activities of the opponent normalization neurons will inhibit the integration neurons, disinhibiting the segregation neurons. These PV neurons will have high spontaneous firing rates to achieve these functions.

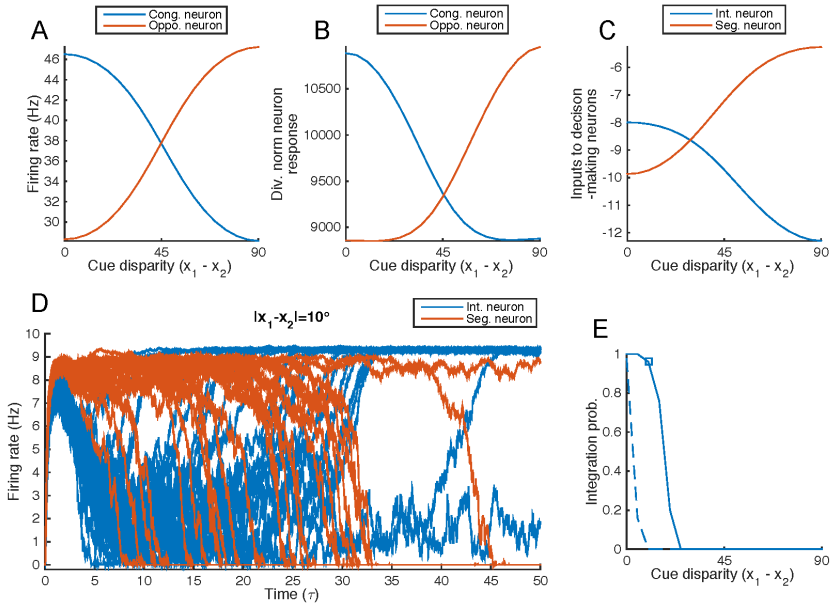

Figure 5: Behavior of the decision circuit as a function of the difference in visual attributes for the two modules. A: With an increase in input disparity between the two modules, the activities of the congruent neurons decrease while that of the opponent neurons increase. The complementary change of congruent and opponent neurons' activities provides a clue to decide whether the inputs are similar or different. B: The activities of normalization neurons affiliated with congruent and opponent neurons. The normalization neuron sums all activities of corresponding neurons, either congruent or opponent neurons. C: The input received by integration and segregation neurons. Apart from receiving the activities of normalization neurons, the integration and segregation neurons also receive summed feedforward and horizontal inputs. D: the activity of the segregation units (red) and that of the integration units (blue). Each trace is for one trial. 25 trials are depicted. Because integration and segregation units compete with each other, thus each red curve coming down is accompanied by each blue curve going up, reflecting an integration decision. E: The probability of integration unit winning is denoted as a square dot in E. A summary of the winning probability of the integration neurons of the decision circuit (solid line), in comparison with the analytical derivation of the targeted optimal Bayes model selection (dotted line).

Fig. 5 shows how the decision circuit behaves as a function of the difference in the visual attributes represented by input disparity between x_1 and x_2 of the two modules. Fig. 5A shows the activities of the congruent neurons decrease while that of the opponent neurons increase with an increase in cue disparity. Fig. 5B shows a similar but sharper decrease for the congruent and the opponent normalization neurons. Fig. 5C shows the input from the congruent and opponent normalization neurons to the segregation and integration neurons respectively. Fig. 5D shows the activity of the segregation units (red) and that of the integration units (blue). Multiple trials are depicted. Each red curve coming down is accompanied by each blue curve going up, indicating integration or segregation. The winning probability of the integration neuron is 0.75. That means that when repeated for a number of trials, integration neurons will win 75% of the time. Fig. 5E and Fig. 6 compare the optimal Bayes model selection probability (dotted line) against the behaviors of the network. We see with increasing input intensity, the integration

probability with cue disparity, $x_1 - x_2$, becomes sharper and also shifts towards to leftward, indicating that two cues will have higher chance to be integrated together when input intensity increases. The current behavior of our network (solid lines in Fig. 6) behaves qualitatively close to the optimal Bayes decision (dashed lines in Fig. 6) but not exactly the same. We are still trying to tune and refine the model to increase this aspect of model's fitness.

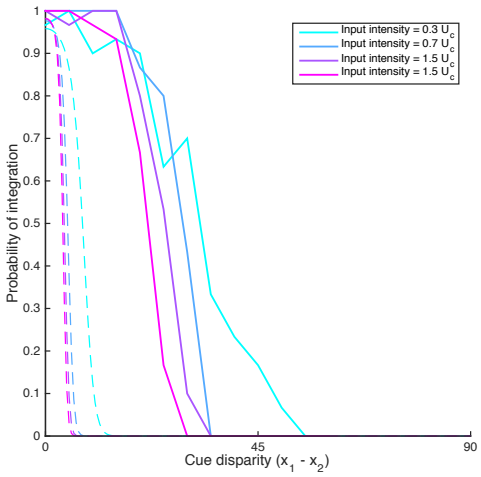

Figure 6: Integration probability with cue disparity and cue intensity. Solid lines are network's performance, while dashed lines are Bayesian prediction. Color denotes the input intensity. With increasing intensity, the integration probability becomes shaper and shifts leftward.

5.2 Fitness of the model in explaining neurophysiological experimental data

In this section, we evaluated the behaviors of our model in three experimental tests to demonstrate that our model can exhibit the statistical learning or the familiarity effect that we observed neurophysiologically (see Section 8 for supplementary neural data). We also evaluated the behaviors of the other models for comparison.

5.2.1 Test 1: Surround modulation of iso and non-iso orientation stimuli

In the first experiment, we evaluated the responses of oriented neurons to three stimulus conditions to evaluate center-surround contextual modulation as in standard neurophysiological experiment. In all these stimulus conditions, the receptive field of the neuron is stimulated by a vertical bar or an input signal corresponding to a vertical bar. We tested three surround stimuli as shown in Fig. 7: (1) center-surround is of the same vertical orientation; (2) center receptive field stimulus is of vertical orientation, the surround is at 45 degrees, slanting to the right; and (3) center RF is vertical, the surround is at 135 degrees, slanting to the left. The first stimulus is used to demonstrate the standard iso-orientation surround suppression. The second stimulus and the third stimulus are for testing the contextual modulation of the non-iso-orientation surround. They are surrogates of the non-Cartesian stimuli that we used in the rodent and monkey neurophysiological experiments such as rays and spirals which we found to show significantly less surround suppression (see Fig. 19 in supplementary section 8.2). We tested these stimuli in the Boltzmann machine, the MGSM, a simple standard baseline model of iso-orientation surround suppression, as well as the proposed model.

To model center-surround contextual modulation, we use module 1 to represent the center