Measures of surround energy.

(a) Cross-validated prediction quality, averaged across neurons, for the standard (orange) and flexible (green) models. Each pair of whisker plots corresponds to a distinct constellation of surround filters or instantiation of the normalization model, as explained below. Boxes denote the 25–75th percentile; whiskers, 10–90th percentile; white line, median; white circle, mean. In all cases, the flexible model out-performed the standard model (p<0.0001). We note that although there was little variation in average performance across different instantiations of surround tuning, we did find that some instantiations performed better than others in each unit. The difference in prediction quality between the best and worst measure was larger than 25% in 63/126 neurons, and larger than 50% in 34/126 of the neurons. However, no definition of the surround consistently performed better than any another. (b) Average MR across all neurons, for images providing weak or strong surround drive (as in main Figure 3b). Conventions as in (a), except white circle indicates geometric mean. In most cases, suppression was not stronger when surround drive was strong vs. weak (p≥0.02 in all cases, except “image-based”, “Lp-norm” and “Free exp” for which p<0.0001).
To test the robustness of our results to other measures of surround drive, we considered the following variations on the model presented in the main manuscript:

(1) Surround strongly tuned for orientation and spatial frequency (denoted “narrow ori, sf”): The surround was composed of filters matched to those representing the RF, including the same orientation preference (which was fixed across images). In total, this surround used 16 filters, 8 locations x 2 for the quadrature pair at each location.

(2) Surround strongly tuned for orientation but not for spatial frequency (denoted “narrow ori, broad sf”): The surround was composed of filters matched to those representing the RF, but with 3 different spatial frequency preferences. In total, this surround used 48 filters, 8 locations x 3 spatial frequencies at each location x 2 for the quadrature pair at each location.

(3) Surround tuned for spatial frequency but not for orientation (denoted “broad ori, narrow sf”): The surround was composed of filters matched to those representing the RF, but with 4 different orientation preferences. In total, this surround used 64 filters, 8 locations x 4 orientations at each location x 2 for the quadrature pair at each location.

(4) Surround broadly tuned for orientation and spatial frequency (denoted “broad ori, sf”): The surround was composed of filters with 4 different orientations and 3 different spatial frequencies. In total, this surround used 192 filters, 8 locations x 4 orientations at each location x 3 spatial frequencies at each location.

(5) Spatially in-homogeneous surround (denoted “location-based”): We also considered the possibility that surround suppression was spatially inhomogeneous (Cavanaugh et al. J Neurophysiol 88, 2547-2556 (2002); Walker et al. J Neurosci 19, 10536-10553 (1999)). We used the filters in configuration (3) above, but allowed the gain to vary across spatial locations. The two locations at the ends of the RF (collinear with the RF-preferred orientation) had a gain \( \gamma_{\text{col}} \); the two locations at the sides (parallel but not collinear) of the RF had a gain \( \gamma_{\text{par}} \); and the two pairs of locations along the diagonals had gains \( \gamma_{\text{diag1}} \) and \( \gamma_{\text{diag2}} \). Thus, the surround activity for a given image, \( \gamma E_s \), is replaced by

\[
\gamma_{\text{col}} E_{s-\text{col}} + \gamma_{\text{par}} E_{s-\text{par}} + \gamma_{\text{diag1}} E_{s-\text{diag1}} + \gamma_{\text{diag2}} E_{s-\text{diag2}}.
\]

(6) Normalization with Lp norm: In the main text we considered a form of the normalization model used previously for surround suppression (Cavanaugh et al. J Neurophysiol 88, 2530-2546 (2002)). We chose this form because it has a minimal number of free parameters. With this choice, the normalization pool computes the L2 norm of the filters (square root of the sum of squared filter outputs, which amounts to image contrast in the filters’ frequency and orientation band) followed by a fixed exponent of 1. In a more general form of the normalization model (Carandini and Heeger Nat Rev Neurosci 13, 51-62 (2011)), the normalization pool computes an Lp norm (p-th root of the sum of p-th power of filters outputs, with p a free parameter rather than fixed to 2), and the numerator and denominator can use different exponents.

To be sure that our results did not rely on the particular way in which combined filter outputs, we fit two additional models using the image-based surround used in the main text: (i) we calculated the drive to the RF and surround using an Lp norm, where p is a free parameter, rather than the L2 norm (denoted “Lp-norm”); (ii) we used the L2 norm but allowed the exponent for the numerator and denominator to be different from each other and different from 1 (denoted “Free exp.”).

Finally, in order to provide a baseline against which to compare the standard normalization models, we also estimated prediction quality for a ‘center model’ without normalization. Responses in this model are defined by the numerator of the standard normalization model (eq. (1) of Methods). For this model, the prediction quality was on average 0.13 (c.i.=[0.06 0.21]), only slightly better than the null model which predicts that the response to each image is equal to the average response across images.
Continuous gating of surround suppression.

(a,b) Distribution of inferred probabilities of homogeneity (abscissa), for the images used in the experiment (a) and a larger database of natural images (N=10,000) used to train the Bayesian model (b). Each image contributes multiple times, with different values depending on the filters used for the inference. Inferred probabilities were usually near zero or 1. We verified that a flexible model in which we used the inferred probability for gating (rather than discretizing the probability to a 0 or 1) did not improve the predictions, presumably because most images had probabilities near 0 or 1. (c) To test whether the gating of the surround was continuous or binary, we conducted an additional experiment, recording 91 neurons in 1 animal using methods identical to those in the main text. We used a subset of the experimental images, which consisted of the 10 most frequently homogeneous images and the 10 most frequently heterogeneous images. We used two manipulations to vary the probabilities in a continuous way. First, we rotated the surround of the images by 15, 30 and 90 degrees with respect to the center, reducing the probability that the image is inferred to be homogeneous. Second, we morphed pairs of images, in which one image had an inferred probability close to 0 and the other close to 1, by taking linear combinations of the images with relative weights of 0.25/0.75, 0.50/0.50, and 0.75/0.25, leading to intermediate values of inferred probability of homogeneity. This panel shows the average NMR (defined in main text Methods) across images and neurons with different levels of inferred probability (abscissa); solid green line, best linear fit (weighted least squares). Suppression strength varies linearly with inferred probability of homogeneity, suggesting that the gating of the surround is continuous, not binary. (d) Black symbols, average surround drive (defined in main text Methods) across images and neurons with different levels of inferred probability (abscissa); solid green line, best linear fit (weighted least squares). There is little dependence of drive on inferred homogeneity, suggesting that the trends in (c) cannot be explained by changes in surround drive. Error bars and dashed lines represent 68% c.i.
Supplementary Figure 3

Comparison of standard and flexible model predictions.

(a) Predicted MR vs. observed MR for two example neurons (rows). Values on the abscissa are repeated in the two columns, to compare the predictions of the standard and flexible models for the same neuron. Blue symbols are for images classified as homogeneous by the Bayesian model; red symbols, heterogeneous. Error bars denote 68% c.i.

(b) Predicted MR vs. observed MR across all images and neurons. Dots represent means, error bars denote 68% c.i. In both the example neurons and the population, the standard model tends to predict stronger suppression than observed (smaller MR values), particularly for heterogeneous images (red symbols in (a)). To be sure that the improved prediction quality of the flexible model was not due simply to muting the surround in a random subset of images, we also verified that a flexible model with random inference about image homogeneity was not better than the standard normalization model (prediction quality 0.5, c.i. [0.42 0.56]; p=0.41 for the comparison to the standard model).
**Supplementary Figure 4**

**Inference about homogeneity measures center-surround similarity.**

The map depicts the probability that visual inputs producing center and surround energies $P_k$, $P_s$ (defined in Supplementary Modeling 2), respectively, are inferred homogeneous by the MGSM model. The thick, black contours separate regions with probability larger vs. smaller than 0.5. (b) Same as (a), but computed using the measure of similarity between center and surround defined in Supplementary Modeling 2 ($\alpha = 7; \beta = 1.7\alpha$). The heuristic (b) provides a reasonable approximation to the full Bayesian inference (a).
Supplementary Figure 5

Reduced surround suppression with images that lack natural structure.

(a) To test the prediction that images lacking higher-order statistical dependencies are heterogeneous, and therefore should not be suppressive, we conducted an additional experiment. We recorded 91 neurons in 1 animal, using methods identical to those in the main text. We created white noise images—which lack statistical dependencies—by choosing the luminance value of each pixel from a Gaussian distribution, with mean set to the mean screen luminance, and standard deviation set to 25% of the pixel range [0 255]. Samples outside that range were clipped. Each pixel was 0.08 x 0.08 degrees in size. We created 10 large (6.7 degrees diameter) noise images, and presented each image both at full size and windowed to a 1 degree diameter. Each line in this panel compares the MR for the preferred grating (left) and white noise images (right) for a neuron. White circle indicates the average MR across all neurons. White noise images result in much weaker suppression than preferred gratings (MR noise = 0.96; MR gratings = 0.26; p<0.0001), consistent with the predictions of our framework. (b) We also tested a second manner for removing higher-order statistical dependencies: we phase-scrambled natural images, which removes higher-order structure while maintaining the amplitude spectrum. For this manipulation, we chose a small subset (N=10) of the experimental images that were most often inferred homogeneous. We left the image inside the RF unaltered, but replaced the part of the image in the surround by its phase-scrambled version (Guo et al. Eur J Neurosci 21, 536-548, 2005). The phase-scrambling was achieved by taking the 2D Fourier transform of the image in polar coordinates, randomly shuffling the phase values, and taking the inverse Fourier transform back to pixel space. Each line in this panel compares the MR for a subset of natural images (N=10; left) with that of the same images after phase scrambling; white circles indicate the average MR across all neurons. Suppression was substantially weaker in the phase-scrambled images (MR phase-scrambled = 0.82; MR natural = 0.34; p<0.0001), consistent with the prediction of our framework.
Supplementary Figure 6

Distribution of phase-sensitivity index.

F1/F0 index was defined as the ratio of the amplitude of the first harmonic to the mean response across four phases (0, 90, 180 and 270), measured using the static grating whose orientation and spatial frequency evoked the strongest response. Values close to 0 indicate weak or no sensitivity to gratings phase (i.e. complex cells), whereas values of 1 or larger indicate phase-sensitive (simple) cells. Most cells showed little phase modulation, consistent with our use of a quadrature pair to represent the RF.
Inference about image homogeneity based on measured neuronal tuning.

(a,b) Same as in Figure 6a,b of the main text, but representing each cell with a filter chosen based on tuning measurements rather than chosen to maximize fit quality. Specifically, we measured each cell’s tuning with static grating stimuli (described in Methods). The orientation preference of the neuron was defined by fitting a Von Mises function to the responses to small gratings presented at 16 orientations. The preferred spatial frequency and size of the RF were defined as the location of the peak of the respective measured tuning curves, for the grating orientation that matched the preferred orientation most closely. The filters were then chosen to match the orientation, spatial frequency and size preference. We included in our analyses all neurons for which the filter outputs agreed well with the orientation and spatial frequency tuning ($R^2 \geq 0.5$; n=41 neurons). (c,d) same as in Figure 7d,e of the main text, but for filters chosen based on measured neuronal tuning. Together, these analyses show that the importance of image homogeneity for explaining suppression does not depend in any way on fitting filters to responses—it is evident even when the neuron’s filters are measured directly with gratings. However, the stronger suppression for homogeneous images was contingent on correct estimation of the filters representing each neuron: the difference between image classes either vanished or changed sign when we used filters with a randomly chosen tuning profile, rather than those based on gratings measurements or fitting (not shown).
Supplementary information

Flexible Gating of Contextual Influences in Natural Vision
R. Coen-Cagli, A. Kohn, O. Schwartz

Supplementary Table 1 ................................................................................................................... 2
Supplementary Modeling ................................................................................................................ 3
  1. Bayesian model and image statistics ................................................................................... 3
  2. Image homogeneity and center-surround similarity ........................................................... 7
Supplementary Table 1

<table>
<thead>
<tr>
<th></th>
<th>Population size</th>
<th>Prediction quality (flexible, standard)</th>
<th>$p$</th>
<th>$MR$ (hom, het) $p$</th>
<th>$NMR$ (hom, het) $p$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Main text</strong></td>
<td>126</td>
<td>0.69, 0.53</td>
<td>&lt;0.0001</td>
<td>0.47, 0.62</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Least phase-sensitive units</td>
<td>63</td>
<td>0.69, 0.51</td>
<td>&lt;0.0001</td>
<td>0.48, 0.64</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Most phase-sensitive units</td>
<td>63</td>
<td>0.69, 0.55</td>
<td>&lt;0.0001</td>
<td>0.47, 0.61</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Filter properties from gratings</td>
<td>41</td>
<td>0.7, 0.51</td>
<td>&lt;0.0001</td>
<td>0.41, 0.62</td>
<td>0.93, 1.53</td>
</tr>
<tr>
<td>Most well-isolated single units</td>
<td>30</td>
<td>0.74, 0.64</td>
<td>0.004</td>
<td>0.43, 0.57</td>
<td>0.002</td>
</tr>
<tr>
<td>Full population</td>
<td>207</td>
<td>0.58, 0.5</td>
<td>0.0001</td>
<td>0.57, 0.69</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

**Supplementary Table 1.** Results were robust to neuronal exclusion criteria, isolation quality, and method for defining model filter properties. The first row contains the values reported in the main text, as a reference. The second and third rows show results for the median split based on the neurons phase sensitivity index, $F_1/F_0$ (Supplementary Figure 6). We obtained similar results in the halves of the data, alleviating concern about the use of a quadrature pair to represent the RF. The fourth row shows results when the filters for the RF center and surround were based on measurements of spatial frequency and orientation tuning made with static gratings (Supplementary Figure 7). The fifth row shows results for only the most well-isolated single units (waveform SNR>2.75, as defined in Kelly et al. J Neurosci 27, 261-264 (2007)). The bottom row includes neurons that were excluded from the analysis of the main text, because they were not strongly surround-modulated. All reported results were thus robust to any of the selection criteria used.
Supplementary Modeling

1. Bayesian model and image statistics

The Bayesian model we used has been described in detail in previous publications (Coen-Cagli et al., Advances in NIPS 369–377, 2009; Coen-Cagli et al., PLoS Comput Biol 8, e1002405, 2012). For related work, see (Schwartz and Simoncelli, Nat Neurosci 4, 819-825, 2001; Schwartz et al., Neural Comput 18, 2680-2718, 2006; Guerrero-Colon et al., IEEE ICIP 565-568, 2008; Hammond and Simoncelli, IEEE Trans Im Proc 17, 2089-2101, 2008; Karklin and Lewicki, Nature 457, 83-86, 2009). Here we provide a brief description of the model, and of the specific implementation used to compare to neuronal responses.

Visual filters and the Bayesian model

Our Bayesian model was a generative model, the Mixture of Gaussian Scale Mixture (MGSM), that captures the characteristic dependencies between the outputs of neighboring filters, when these are applied to natural images (Coen-Cagli et al., Advances in NIPS 369 – 377, 2009; Schwartz et al., Neural Comput 18, 2680-2718, 2006; Wainwright et al., Appl Comput Harmon Anal 11, 89-123, 2001). In our case, the MGSM considers the outputs of two groups of visual filters, representing the RF and surround of a neuron. We derived the V1-like filters from a steerable pyramid (Simoncelli et al., IEEE Trans Inf Theory 38, 587-607, 1992; Portilla and Simoncelli, Int J Comp Vis 40, 49-70, 2000) and defined the filter outputs—$k$ and $s$—as the dot product between an image and the filters. When the filters are applied to homogeneous images, their outputs are dependent, because of global properties of the image. In this case, the filter outputs $k$ and $s$ are generated in the MGSM by multiplying local Gaussian variables by the same positive, scalar random variable denoted by $\nu$ (termed the mixer). The local Gaussian variables—$K$ and $\sigma$—with the same dimensionality as $k$ and $s$—capture the strength of local image features in the RF and surround, respectively. Thus, the RF and surround filter outputs are given by:

\[
\begin{align*}
{k} &= \nu K \\
{s} &= \nu \sigma
\end{align*}
\]

(Eq. S1)

While the dependencies between the elements of $K$ and $\sigma$ are fully described by a covariance matrix $C_{(K\sigma)}$, the joint probability distribution of the elements of $k$ and $s$ also have a variance dependence, because the multiplication by the common mixer $\nu$ scales both variances similarly. This variance (“bowtie”) dependence is a hallmark feature of applying V1-like filters to natural images (Schwartz and Simoncelli, Nat Neurosci 4, 819-825, 2001; Wainwright et al., Appl Comput Harmon Anal 11, 89-123, 2001).
When the filters are applied to heterogeneous images, the filters’ outputs are independent. In this case, the filter outputs are generated as follows:

(Eq. S2)

\[
\begin{align*}
    k &= \nu \kappa \\
    s &= u \sigma 
\end{align*}
\]

where \( u \) is a mixer variable independent of \( \nu \), and \( \kappa \) and \( \sigma \) are described by two separate covariance matrices \( C_(\kappa) \) and \( C_(\sigma) \).

**Inference in the Bayesian model and divisive surround normalization**

To specify the relation between the Bayesian model and the response of a neuron to an image, we assumed that V1 neurons aim to compute an optimal estimate of the local content of the image inside the RF, and thus to remove redundancy with the representation of nearby filters. This assumption is steeped in a rich literature that has argued that cortex tries to achieve an efficient representation by reducing redundancy (Attneave, Psychol Rev 61, 183, 1954; Barlow, in Sensory Communication, 217-234, MIT Press, 1961; Dayan and Abbott, Theoretical neuroscience, MIT Press, 1999), and that has shown that many aspects of cortical responses can be explained by this simple computational goal (Schwartz and Simoncelli, Nat Neurosci 4, 819-825, 2001; Simoncelli and Olshausen, Annu Rev Neurosci 24, 1193-1216, 2001; Olshausen and Field, Nature 381, 607-609, 1996; Bell and Sejnowski, Vision Res 37, 3327-3338, 1997; Simoncelli, Curr Opin Neurobiol 13, 144-149, 2003; Zhaoping, Network 17, 301-334, 2006).

In the context of the MGSM, the local feature in the target neuron RF (defined, for instance, by a filter of preferred orientation \( \theta \) and spatial frequency \( \omega \)) is represented by the value of the underlying Gaussian variable \( (K_{\theta,\omega}) \). Intuitively, given the observed sensory inputs \( (k \) and \( s) \), one could compute \( K_{\theta,\omega} \) by inverting (Eq. S1) or (Eq. S2), which amounts to the division of \( k \) by \( V \). However, whether the outputs are dependent (i.e. (Eq. S1) applies) or independent (i.e. (Eq. S2) applies) for a given image is not known, and \( V \) is also not observed directly. Instead, according to Bayesian inference, an estimate \( K_{\theta,\omega} \) can be obtained by integrating out those unobserved variables (for the full analytical derivation see Coen-Cagli et al., Advances in NIPS 369–377, 2009; Schwartz et al., Neural Comput 18, 2680-2718, 2006).

This operation, termed marginalization, is a general Bayesian prescription to ignore nuisance variables, and has recently been linked to divisive normalization in other contexts as well (Beck et al. *J Neurosci* 31, 15310-15319, 2011).
If we introduce a binary variable \( \xi \) and denote the inferred probability that the filters are dependent or independent as \( p(\xi = \xi_1 \mid k, s) \) and \( p(\xi = \xi_0 \mid k, s) \), respectively, then the estimate of \( \kappa_{\theta, \omega} \) is given by:

(Eq. S3) \[
\kappa_{\theta, \omega} \approx p(\xi = \xi_1 \mid k, s) \frac{|k_{\theta, \omega}|}{\sqrt{\lambda_{(\kappa \tau)}}} + p(\xi = \xi_0 \mid k, s) \frac{|k_{\theta, \omega}|}{\sqrt{\lambda_{(\kappa)}}}
\]
where:

(Eq. S4) \[
\lambda_{(\kappa \tau)} = \sqrt{(k, s) C_{(\kappa \tau)}^{-1} (k, s)^\top}
\]

(Eq. S5) \[
\lambda_{(\kappa \tau)} = \sqrt{k C_{(\kappa \tau)}^{-1} k^\top} = \sqrt{\sum_{i,j=1}^{n_k} (C_{(\kappa \tau)})_{i,j} k_i k_j}
\]

The term \( p(\xi = \xi_1 \mid k, s) \) in (Eq. S3) represents the inferred posterior probability that the filter outputs are statistically dependent. This probability can also be computed analytically (Coen-Cagli et al., Advances in NIPS 369–377, 2009; Schwartz et al., Neural Comput 18, 2680-2718, 2006):

(Eq. S6) \[
p(\xi = \xi_1 \mid k, s) \propto p(\xi = \xi_1) p(k, s \mid \xi = \xi_1)
\]

where \( n = n_k + n_s \) is the total number of filters; \( B \) is the type-2 modified Bessel function; and

\( p(\xi = \xi_1) \) is the prior probability that the outputs are dependent, which we obtained by training the MGSM on a database of natural images (see Supplementary Methods 3). The inferred probability that the inputs are independent \( p(\xi = \xi_0 \mid k, s) \) has a similar expression (Coen-Cagli et al., PLoS Comput Biol 8, e1002405, 2012), but involving the covariances \( C_{(\kappa)}, C_{(\sigma)} \).

When the posterior probability that the RF and surround filter outputs are dependent is large (i.e. \( p(\xi = \xi_1 \mid k, s) \approx 1 \) for a homogeneous image), the estimate of the local component, \( \kappa_{\theta, \omega} \), is proportional
to the ratio between the observed response of the RF filter \( \hat{k}_{\theta,\omega} \) and the normalization signal computed from the responses of all other filters \( \hat{\lambda}_{\text{c.s.}} \), a generalized measure of center-surround energy, as shown in (Eq. S4). That is, the estimate involves divisive normalization from the surround. The generalized measure of center and surround contrast energy reduces to a classical description of normalization (Heeger, Vis Neurosci 9, 181-197, 1992) when the covariance matrix is diagonal (i.e. when the dependency between \( k \) and \( s \) is entirely explained by the shared mixer variable, with no linear correlations between their outputs). Critically, when the probability is small (i.e. \( p(\xi = \xi_1 | k, s) \approx 0 \) for a heterogeneous image) the estimate does not involve surround normalization. Hence the prediction is that surround suppression is muted for heterogeneous images.

**Training the Bayesian model**

In addition to the properties of the center and surround filters, the Bayesian model (MGSM) uses several parameters that are learned from natural images—namely, the prior probability that the outputs are dependent \( p(\xi = \xi_1) \), and the covariance matrices for the two mixture components \( C_{(\text{c.s.})}, C_{(\kappa)}, C_{(\sigma)} \). The values of these parameters determine whether a particular image will be classified as homogeneous or heterogeneous, according to (Eq. S6).

For each set of filters, we trained the Bayesian model on image patches randomly selected from a database commonly used for image compression benchmarks (http://neuroscience.aecom.yu.edu/labs/schwartzlab/code/standard.zip). We computed the outputs of the visual filters to the patches, and searched for the parameter values that maximized the likelihood of the filters’ outputs, namely \( p(k, s | \text{parameters}) \). To find the maximum likelihood parameters we used an iterative procedure based on the Expectation Maximization (EM) algorithm, as it is standard for mixture models (Coen-Cagli et al., Advances in NIPS 369–377, 2009; Meng and Rubin, Biometrika 80, 267-278, 1993). The training was unsupervised: we did not pre-specify which patches in the training set were homogeneous or heterogeneous, but rather let the model infer them. Once we had defined the optimal parameters with the standard database, we froze them before classifying the experimental images.
2. **Image homogeneity and center-surround similarity**

We defined image homogeneity formally, as the result of inference in an MGSM model of natural images (Eq. S6). Intuitively, the inference reflects the statistical similarity of signals in the RF and surround. Here we specify this intuition further, in order to clarify how image homogeneity defines a measure of center-surround similarity and what sort of structure in images leads to their being classified as homogeneous.

We study the properties of the ratio between the probability that an image with filters outputs \((k,s)\) (recall that \(k\) represents the outputs of a set of filters in the RF center with different orientations, and \(s\) a set of filters with fixed orientation and different positions in the surround, as explained in the Methods) is homogeneous vs. heterogeneous:

\[
M = \frac{p(\xi = \xi_0 | k, s)}{p(\xi = \xi_0 | k, s)} = \frac{p(\xi = \xi_0)}{p(\xi = \xi_0)} \left(\frac{(2\pi)^{n_k} \det(C_{(k)})}{(2\pi)^{n_s} \det(C_{(s)})}\right).
\]

(Eq. S7)

\[
B\left(1 - \frac{n_k}{2}; \lambda_{(k)}\right) B\left(1 - \frac{n_s}{2}; \lambda_{(s)}\right)
\]

We would like to understand which patterns of filters outputs lead to higher ratios, corresponding to higher inferred probability that the image is homogeneous.

A heuristic approximation to the inference of homogeneity, for a simplified case

We first consider the following special case, in which there are no correlations between filters (i.e., diagonal covariance matrices) and homogeneous and heterogeneous images are equally probable a priori. Specifically:

\[
\begin{align*}
& p(\xi = \xi_0) = p(\xi = \xi_0) = 0.5 \\
& C_{(k)} = \mathbf{I}_n \\
& C_{(s)} = \mathbf{I}_n \\
& n_k = n_s = 2m \\
& n = 4m
\end{align*}
\]

from which it follows (by substitution in (Eq. S4) and (Eq. S5)) that:
\[
\begin{align*}
\lambda_{(k)} &= \sqrt{\sum_{i=1}^{2m} k_i^2} = \sqrt{P_k} \\
\lambda_{(\sigma)} &= \sqrt{\sum_{i=1}^{2m} s_i^2} = \sqrt{P_s} \\
\lambda_{(k\sigma)} &= \sqrt{\sum_{i=1}^{2m} k_i^2 + \sum_{i=1}^{2m} s_i^2} = \sqrt{P_k + P_s}
\end{align*}
\]

This simplifies (Eq. S7) greatly, to:

(Eq. S8) \[ M = \frac{\mathcal{B}(2m-1; \sqrt{P_k + P_s})}{\mathcal{B}(m-1; \sqrt{P_k}) \mathcal{B}(m-1; \sqrt{P_s})} \cdot \frac{(P_k \cdot P_s)^{\frac{m-1}{2}}}{(P_k + P_s)^{\frac{2m-1}{2}}} \]

We can now verify if the intuition is correct that when the drive to the RF and surround are similar, images are inferred homogeneous (i.e. \( M \) takes on large values). In this case \( P_k = P_s \equiv P \) and therefore

(Eq. S9) \[ M = \frac{\mathcal{B}(2m-1; \sqrt{2P})}{\mathcal{B}(m-1; \sqrt{P})^2} \cdot \frac{1}{(2)^{\frac{m-1}{2}}} \cdot \frac{1}{P^\frac{1}{2}} \]

When the inputs are weak, i.e. \( P \to 0 \), we can use the known limiting form of the Bessel function for small arguments (Abramowitz and Stegun, Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables (1972), p.375):

(Eq. S10) \[ \mathcal{B}(n; x) \approx \frac{1}{2} \Gamma(n) \left( \frac{2}{x} \right)^n \]

and therefore

(Eq. S11) \[ M \approx 2^{\frac{7}{2}} \cdot \frac{\Gamma(2m-1)}{\Gamma(m-1)^2} \cdot \frac{1}{P} \]

thus showing that \( M \) diverges for weak and similar RF and surround drive. Similarly, for strong inputs that are similar in the RF and surround, i.e. \( P \gg 1 \), we can take the first order asymptotic expansion of the Bessel function (Abramowitz and Stegun, Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables (1972), p.377), and find that:

(Eq. S12) \[ M \approx \frac{1}{\pi 2^{m-2}} \cdot e^{2RF} \]

again showing that \( M \) diverges for large and similar center and surround energies.
We can also verify the intuition that when the energies in the center and surround are different, images are inferred heterogeneous (i.e. $M$ takes on small values). We consider first the case that $P_s \gg P_k$, or equivalently $P_k \to 0, P_s \to \infty$. Under this circumstance we find:

(Eq. S13) \[ M \approx \frac{B(2m-1; \sqrt{P_k})}{B(m-1; \sqrt{P_k})} \frac{P_k^{m-1}}{P_s^{2m-1}} \]

and using the limiting form and asymptotic expansion of the Bessels, for $P_k \to 0, P_s \to \infty$ respectively, we find:

(Eq. S14) \[ M \approx \frac{(P_k)^{m-1}}{(P_s)^{m}} \]

This ratio vanishes as the numerator goes to zero and denominator to infinity, thus confirming that images with strong surround and weak center energy are inferred heterogeneous. A similar result holds for strong RF input paired with weak drive to the surround.

The behavior for intermediate values of $P_k, P_s$ cannot be described as easily analytically. However, we can compute $M$ exactly via (Eq. S8) for a range of values of $P_k, P_s$ commonly observed on natural images. In Supplementary Figure 4a we assumed $m=8$ (i.e. 16 filters, or 8 quadrature pairs, in the center and 16 in the surround), and plot the values of $M/(M+1)$, which is identical to $p(\xi = \xi^* | k, s)$—the posterior probability that the image is homogeneous. As expected, the largest values are found closer to the diagonal, when $P_k, P_s$ are similar. This is true not only in the asymptotic regimes but also for intermediate energies. However, to achieve a high probability of inferred homogeneity, there is a more stringent requirement that $P_k, P_s$ are similar at low values of $P_k, P_s$ than at high values. Thus, the requirement of similar drive to the center and surround has a strong intensity-dependence. Another way to think about this result is in terms of the so called Bayesian Occam’s razor: when the evidence is weak (e.g. a low contrast image, corresponding to low values of $P_k, P_s$), stick to the simpler model (the center and surround signals are independent) rather than the more complex one (center and surround signals are dependent).

We then searched for a simpler, intuitive similarity measure that could approximate the behavior of the full inference. We found that the measure $S = \frac{(P_k \cdot P_s)^x}{(P_k + P_s)^y}$, based on the rightmost term in (Eq. S8), did reasonably well: it increased both when the energies were more similar and when they were larger.
(i.e. when the product dominates over the sum). Importantly, simpler measures, based on either the difference between $P_k$ and $P_s$, or their log-ratio, could not reproduce the characteristic intensity-dependence of the full inference. Thus, simpler measures of similarity—such as those suggested by previous physiological studies—fail to capture the behavior of the inference of homogeneity.

**Supplementary Figure 4b** illustrates the probability corresponding to the measure $S$ (namely $S/(S+1)$, which is analogous to $M/(M+1)$ plotted in **Supplementary Figure 4a**). We searched manually for values of the exponents in $S$, and found that the qualitative behavior of $p(\xi = \xi_1 | k, s)$ could be approximated reasonably well for a large range of values, provided that $\beta \approx 1.7\alpha$.

In summary, this analysis confirms the intuition that MGSM inference about homogeneity defines a measure of center-surround similarity. Such similarity is input strength-dependent, and can be approximated as a product-over-sum ratio with different exponents.

**Inference with realistic covariance structure**

So far we have considered the case where the covariances are identity matrices. However, in the MGSM the covariances are optimized to capture structure in natural images, and in practice they are far from diagonal (Coen-Cagli et al. PLoS Comput Biol 8, e1002405 (2012)). How do non-diagonal covariances affect the inference? As shown in (Eq. S4) and (Eq. S5), the entries of the inverse covariances act as weights in the sum of filter energy terms, therefore determining the values of $\lambda_{(k)}$, $\lambda_{(s)}$, $\lambda_{(k\sigma)}$. As shown in (Eq. S7), the values of these $\lambda$ terms directly influence the inference of homogeneity: when $\lambda_{(k\sigma)}$ is small and the product of $\lambda_{(k)}$, $\lambda_{(s)}$ is large, the image is more likely to be homogeneous; and conversely when $\lambda_{(k\sigma)}$ is large and the product of $\lambda_{(k)}$, $\lambda_{(s)}$ is small.

To gain some intuition, consider $\lambda_{(k\sigma)}$. Its value is small when the term $(k, s)C_{(k\sigma)}^{-1}(k, s)^\top$ is large (and therefore $(k, s)C_{(k\sigma)}(k, s)^\top$ is small). The term $(k, s)C_{(k\sigma)}(k, s)^\top$ is large when $(k, s)$ is a pattern that matches structure in the covariance. For instance, for natural images the variance and covariance of surround filters collinear with the RF (i.e. similarly oriented, and placed at the ends of the RF) are much larger than those of other surround filters (e.g. similarly oriented, but placed at the sides of the RF; Coen-Cagli et al. PLoS Comput Biol 8, e1002405 (2012)), due to the predominance of long lines and edges in scenes. Therefore, the image of a long stripe passing through the RF and collinear surround filters will produce a large value of $(k, s)C_{(k\sigma)}(k, s)^\top$, and therefore a small value of $\lambda_{(k\sigma)}$, leading to a higher probability of inferred homogeneity. An image containing similarly oriented structure in the RF
but no collinearity, will lead to a smaller value of $(k, s)C_{(\sigma \sigma^T)}(k, s)^\top$, and therefore a lower probability of inferred homogeneity.

More precisely, let us consider the eigen-decomposition of the covariance $C_{(\sigma \sigma^T)} = \sum_i h_i (v_i \cdot v_i^\top)$ and its inverse $C_{(\sigma \sigma^T)}^{-1} = \sum_i h_i^{-1} (v_i \cdot v_i^\top)^{-1}$, where $v_i, h_i$ are the orthonormal eigenvectors of the covariance and corresponding eigenvalues, respectively. If the input pattern $(k, s)$ is an eigenvector $v_i$ with large eigenvalue $h_i$ (i.e., it matches the dominant structure in the covariance), then $\lambda_{(\sigma \sigma^T)}$ amounts to $\sqrt{1/h_i}$, a small number. This leads to a higher probability of inferred homogeneity. Conversely, if the input pattern $(k, s)$ is an eigenvector with small eigenvalue, then $\lambda_{(\sigma \sigma^T)}$ is large, leading to a lower probability of inferred homogeneity.

In summary, visual inputs that resemble the dominant structure (the largest eigenvectors) in $C_{(\alpha \sigma^T)}$, and therefore differ from the dominant structure in $C_{(\alpha)}, C_{(\sigma)}$, are assigned high probability of being homogeneous. Therefore the full inference about homogeneity in the MGSM amounts to a generalized measure of center-surround similarity, which is dependent on the overall oriented energy in the image, and which uses a metric defined by the structure of the covariance matrices learned from natural images.