Supplementary discussion 1: Most excitatory and suppressive stimuli for model neurons

The model allows us to determine, for each model neuron, the set of most excitatory and suppressive features. First, we compute the covariance given by turning on only one neuron ($y_j = 1$) and leaving the rest at 0,

$$\mathbf{C} = \exp\left(\sum_k w_{jk}\mathbf{b}_k\mathbf{b}_k^T\right) . \tag{S1}$$

This fully specifies the distribution of images encoded by neuron j, and accounts for all the contributions of individual features $\mathbf{b}_k$. Next, we compute the eigenvector decomposition of this matrix. The set of eigenvectors and eigenvalues describes how this distribution differs from the canonical distribution (whose covariance is the identity matrix and whose eigenvalues are all equal to 1). Eigenvectors with the largest eigenvalues correspond to directions in image space that are most expanded; these are image features that maximally excite the neuron (y_j is positive and large). Eigenvectors associated with the smallest eigenvalues represent directions that are most contracted; the presence of these image features suppresses the neuron. This is illustrated schematically in Fig. S1, which also shows the most excitatory and suppressive features for the neuron analyzed in Fig. 3.

Supplementary discussion 2: Relationship to spike-triggered covariance

Eigenvector analysis of model parameters is closely related to spike-triggered covariance (STC), a technique used to characterize response properties of non-linear sensory neurons [1,2]. In this analysis, the covariance of stimuli that elicited a spike is compared to the covariance of the entire stimulus ensemble. Eigenvector decomposition is then used to identify directions in stimulus space (e.g. image features) along which the covariances most differ; these are the stimulus dimensions that most affect neural response. In the case of visual neurons, eigenvectors with the largest eigenvalues correspond excitatory image features to which the neuron is maximally sensitive, while those with the smallest reveal the most suppressive features. As in the eigenvector analysis of the

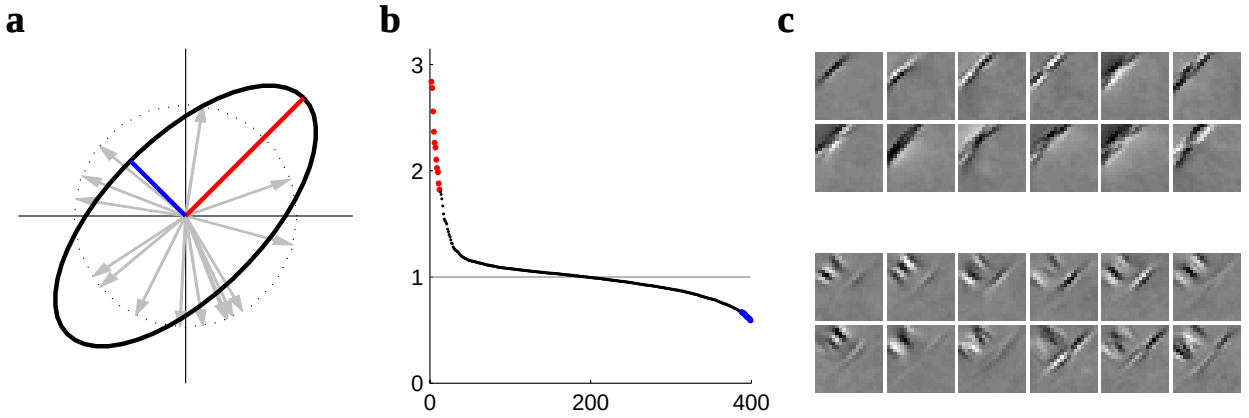

Figure S1: **a**. A schematic of one model neuron's effect on the encoded distribution. The neuron uses the underlying image features (gray arrows) to transform the canonical distribution (dotted circle) into a different distribution (black ellipse). The effect of the neuron on the distribution is given by the eigenvector decomposition of the resulting covariance matrix (see text). The most expanded and most contracted directions correspond to the largest and smallest (respectively) eigenvalues (red and blue lines). **b**. For the model trained on 20×20 images, the full set of 400 eigenvalues describes the scale of all directions in image space. Here we plot the eigenvalues of the model neuron analyzed in Fig. 3. **c**. Eigenvectors associated with the largest 12 (top) and the smallest 12 (bottom) eigenvalues, drawn in image form. The corresponding extreme eigenvalues are highlighted in color in **b**.

proposed model, this method characterizes a *distribution* of images and identifies entire subspaces of inputs over which the neural response is largely invariant. These subspaces do not necessarily correspond to anatomically distinct inputs to the cell (specific presynaptic neurons).

In addition to the eigenvector decomposition of the covariance, it is also possible to directly measure STC on the model *responses* (the MAP estimates $\hat{\mathbf{y}}$). The two methods are not equivalent, since the distribution of these estimates is not that same as the distribution $p(\mathbf{y})$ assumed by the model. In practice, however, we find that probing model neurons with white noise and computing the STC on $\hat{\mathbf{y}}$ yields image features that are nearly identical to the eigenvectors computed from model parameters.

In Fig. S2, we compare eigenvector analysis of three model neurons to data from V1 complex cells[3]. STC, when applied to complex cells, recovers a variable number of excitatory and suppressive image features, which are typically oriented and localized in space[3,4]. For some neurons, only excitatory features are recovered, and the suppressive effects are usually weaker than the excitatory, although they are nevertheless important for predicting a neuron's response [4]. The proposed model predicts subfields that are qualitatively similar to these measurements. The dominant com-

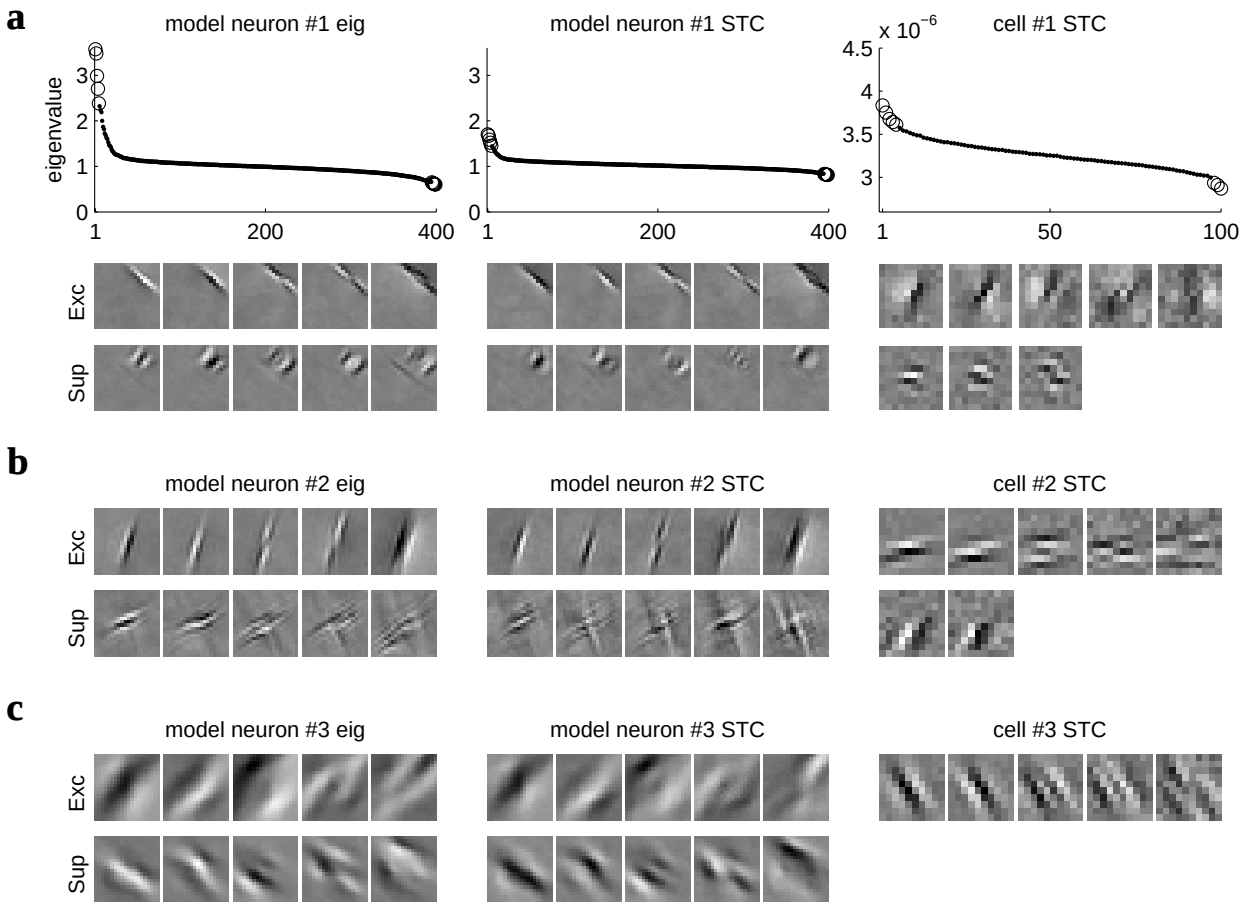

Figure S2: **Spike-triggered characterization of model neurons reveals functional subunits similar to those found in V1 complex cells.** **a**. Eigenvalues and eigenvectors for one model neuron, as revealed by the eigenvector analysis of model parameters (first column) and STC of model responses to white noise (second column). The top row of image patches shows the most excitatory dimensions, the bottom, the most suppressive. These correspond to eigenvectors with the largest and smallest eigenvalues (plotted as open circles). Subunits of similar shape have been measured in complex cells in V1 (3rd column, data reproduced from a physiological experiment[3]). **b**,**c**. Principal eigenvectors for two additional model neurons and complex cells with similar properties, analyzed as in (**a**).

ponents of the excitatory subspaces are oriented features at different phases and positions. These are thought to underlie response invariance to small physical transformations and correspond to functional subunits in the classical model of complex cells[5], but unlike the classical model, both physiology and model predictions suggest an integration of more than two image features. Suppressive effects are typically weaker and comprised of orthogonal image features (Fig. S2a). The model population also includes neurons with non-orthogonal suppression (Fig. S2b). Finally, some neurons broadly integrate oriented features at multiple scales and across a large portion of the receptive field (Fig. S2c).

One discrepancy between the theoretical predictions and complex cell properties is that in the model, suppression is invariably a strong effect, whereas some neurons in V1 appear to have only excitatory subunits. This could be an artifact of using noisy measurements to assess the significance of eigenvalues and eigenvectors (STC of model responses also produces weaker and noisier suppressive estimates). Another possible cause is that the current form of the model assumes a symmetric prior distribution over neural activity (positive and negative values of y_j are equally likely). This tends to favor solutions with balanced excitatory and suppressive effects, and also groups different types of statistical structure that might be better encoded with separate "on" and "off" channels.

Supplementary discussion 3: Types of neurons in the population

In order to quantify the properties of the learned population of neurons, we performed the physiological analysis shown in Fig. 3 for all model neurons. A large subset of the population (42 of 150 neurons) was tuned for orientation; the majority of these (n=35) were insensitive to the phase of the test grating. Many of the orientation-tuned neurons also exhibited the effects shown in Fig. 5: 90% (n=38) were significantly suppressed by an orthogonal masking grating and in 67% (n=28) the response was weaker when an orthogonal annulus was placed in the surround. The median orientation bandwidth of orientation-selective neurons was 46°. (See supplemental methods 3 for details of these tests.) Many of these response types have been observed in V1 cells, but it is difficult to make more detailed comparisons between the model and neural populations. This is due in part to the dependence of the model population composition on assumptions such as the form of the prior or the size of the network. Also, it is problematic to make detailed comparisons to physiological data at the population level, because until recently it has not been possible to systematically characterize non-linear response properties of visual neurons. Data-driven techniques such as spike-triggered covariance analysis as discussed above promise to fill this gap.

To identify groups of neurons in the model with similar function, we performed cluster analysis (Fig. S3) with which we examined how the neurons integrate information from linear features $\mathbf{b}_k$. Specifically, we derived a simple parameterization of the features (their location, orientation, and the dominant spatial frequency), and computed which of these could best account for the values

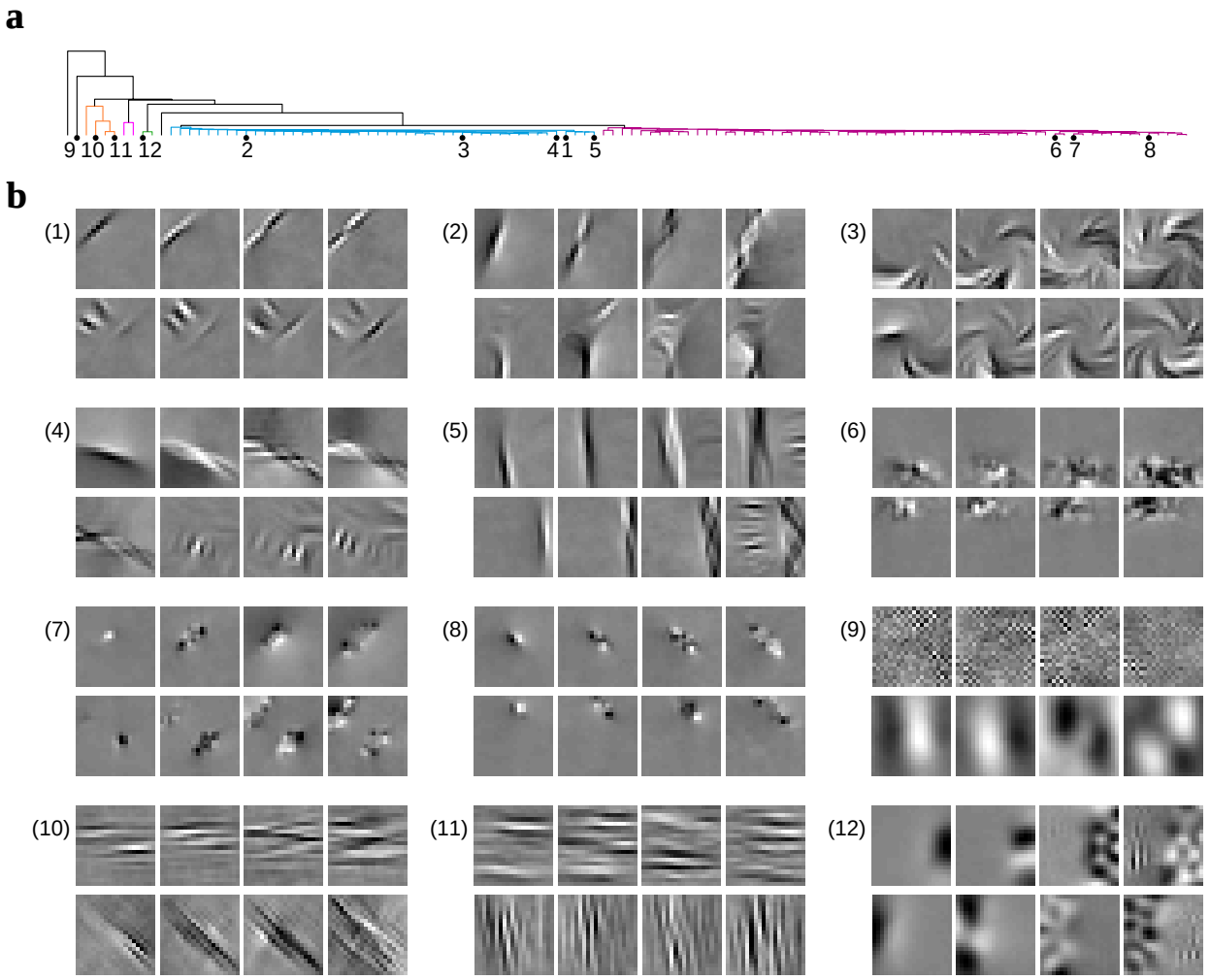

Figure S3: **Image distributions are encoded by a diverse population of neurons. a**. 120 most active (out of a total of 150) neurons were hierarchically clustered according to the different aspects of image structure they encode (see text for details). The clustering reveals two large categories of neurons, as well as some specialized neurons. Subtrees are distinguished in color for visibility. **b**. To obtain a concise description of each neuron, we identified its most activating and most suppressive image features (see Supplementary Discussion 1). Here, for twelve model neurons that are representative of the learned population, we show four excitatory (top row of each panel) and four suppressive (bottom row) image features. Numbers indicate the neuron's position in the dendrogram in (**a**). Neuron (1) was analyzed in Fig. 3.

of a neuron's weights w_{jk}. For example, the neuron in Fig. 3 is sensitive to oriented and localized structure, and we expect its weights to the underlying image features (i.e. the colors in Fig. 3b) to be explained best by the location and orientation of features $\mathbf{b}_k$ (in fact, these parameters account for 93% of the variance of its w_{jk}'s). For each neuron, we computed a vector that indicated how much the feature parameters (as well as all their combinations) contributed to explaining the neuron's weights, and then used standard hierarchical clustering methods (single linkage algorithm) to produce a dendrogram from these vectors.

This analysis revealed two large groups and a small number of specialized units; the population exhibits a range of properties observed in cortical visual cells. One large set is characterized by localized, oriented excitatory features (e.g. the neuron in Fig. 3, also shown in the first panel of Fig. S3b). Most exhibit the inhibitory cross-orientation and surround regions described in Fig. 3 associated with orientation-selective V1 and V2 neurons, while encoding a variety of image types, some with curvature or more complex patterns (1-5). Another large set of neurons is employed by the model to indicate localized contrast (energy) in the stimulus (6-8). Individually, each of these specifies only coarsely the location of contrast energy in the stimulus (and corresponds to a broad set of image distributions), but their joint activity acts as a set of constraints that input images must satisfy to belong to the encoded distribution. Although cortical neurons have not been analyzed in a framework that could identify such a code, localized contrast subfields are consistent with observations that many cortical neurons are sensitive to second-order (energy) patterns in the image[6,7].

Among the neurons in the model, some analyze the spatial frequency content of the image (e.g. neuron 9). When neuron (9) is active, the input image is inferred to come from a set of images with given frequency statistics. Note that each neural activity in the model can be both positive and negative; positive activity here signals high frequency (fine) image structure, while negative activity signals low spatial frequency (coarse) structure. This neuron does not signal anything about the spatial localization of structure in the image or its dominant orientation, and images that activate it can be quite different, as long as they satisfy the spatial frequency constraints. Other neurons in the population convey global orientation structure (10,11) but are insensitive to the spatial frequency content of the image. Such encoding properties have been observed in V4 neurons, some of which are narrowly tuned for orientation, while others encode frequency information[8]. Other neurons in the model indicate contrast in spatial frequencies across image locations (12), signaling a boundary of textures characterized by their statistical properties. Studies of texture boundary coding in visual cortex have been limited to simple synthetic stimuli[9–13]; these results suggest ways to use more complex textures, defined in a statistical framework, to analyze neural responses.

Note that model predictions are limited to spatial patterns in images because the model was not

trained on temporally varying data and thus cannot capture temporal statistics of natural scenes. However, a similar approach can be applied to image sequences and might explain temporal properties of neural responses in the visual cortex.

Supplementary discussion 4: Feed-forward computation of model neuron activity

In order to compute the neural representation vector $\mathbf{y}$, we must iteratively solve for the most likely neural code given an input image. Nevertheless, it is possible to approximate this computation with a single feed-forward step in which each image is first projected onto the set of vectors $\mathbf{b}_k$. Each neuron takes the squared output of these projections, subtracts 1 along each dimension (a thresholding operation that effectively compares the given image to the canonical distribution of equal variances in all directions), and correlates the resulting pattern against its weights w_{jk}. The gradient used for computing the *maximum a posteriori* values $\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} p(\mathbf{y}|\mathbf{x}, \mathbf{b}_k w_{jk})$ incorporates these computational steps:

$$\frac{d \log p(\mathbf{y}|\mathbf{x}, \{\mathbf{b}_k\}, \{w_{jk}\})}{dy_j} \propto \sum_k w_{jk} \left[\left(\mathbf{b}_k^T \mathbf{x} \right)^2 - 1 + \ldots \right] + \psi'(y_j) \,, \tag{S2}$$

where $\psi'(y_j) = d \log p(y_j)/dy_j$ places the sparse prior on neural activity (higher order terms in the gradient have been omitted).

This is a generalization of the classical energy model of complex cells[5], in which the output of two linear filters is squared and added. Here, a larger number of features are integrated, some excite while others inhibit the neuron, and rather than raw activation, energy relative to a canonical pattern (of equal variation) is encoded. The neural code computed using this approximation is close to the optimal solution, but this feed-forward computation does not account for competition among neurons to achieve best encoding.

Supplementary discussion 5: Relationship to previous models

We have previously published a description of a related statistical model[14,15]. This model was

derived as a hierarchical extension of earlier linear models[16,17] and thus is easier to place in the context of theoretical models of visual processing. The model described here is a generalization of this work. Whereas our previous model learned statistical structures in the output of linear features (specifically, the magnitudes of their variation), the current model can capture changing correlations among the linear features as well. This makes it a more flexible model of probability distributions. The model handles overcomplete lower-level representations (vectors $\mathbf{b}_k$) in a more natural framework and describes each local image distribution as a multi-variate Gaussian, which is a relatively simple, yet rich model of statistical regularities. A quantitative comparison of these models, using coding cost evaluated on a held-out set of image patches, confirmed that the proposed model gives a better statistical description of natural images (data not shown).

The model proposed here learns a distributed code for probability distributions by defining a hierarchical statistical model in which the input image is represented at different levels of abstraction, first by a set of linear features, then by the neural activities that represent the most likely density containing the input. We have used a multivariate Gaussian whose covariance is a function of the neural activity. This has the advantage that the model can in principle describe arbitrary correlation patterns in features while still being mathematically tractable. While the experimental data suggest that the visual system uses representations that are in some ways similar to those of the model, it is possible that there are other models that better describe the types of statistical structure in natural images. Other hierarchical models for unsupervised learning of statistical structure in images have been proposed[18–20]. It is possible those learn similar structures and make interesting predictions about cortical representations, though they have not been analyzed in this framework.

Other models have explored neural coding of probability distributions, for example as a means of representing uncertainty[21], or optimally integrating multiple sources of information[22,23]. In our model, however, encoding probability distributions serves a different purpose, allowing model neurons to generalize across inherent variability in natural scenes. This computational goal is distinct from the issue of noise in perception and leads to a number of novel predictions for representing visual information in the cortex.

Supplementary methods 1: Model details

Given the representation $\mathbf{y}$, the image $\mathbf{x}$ is described by a multi-variate Gaussian distribution with zero mean,

$$p(\mathbf{x}|\mathbf{y}) = \frac{1}{(2\pi)^{N/2}|\mathbf{C}|^{1/2}} \exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{C}^{-1}\mathbf{x}\right), \tag{S3}$$

where N is the dimensionality of the data and $|\mathbf{C}|$ is the absolute value of the determinant of the covariance matrix $\mathbf{C}$.

The covariance matrix is defined in terms of neural activity using the matrix logarithm transformation,

$$\log \mathbf{C} = \sum_{jk} y_j w_{jk} \mathbf{b}_k \mathbf{b}_k^T . \tag{S4}$$

We fixed the norm of vectors $\mathbf{b}_k$ to 1, as the weights can absorb any scaling.

To write the model likelihood (the function we are interested in maximizing) in terms of simple mathematical operations, we use the following relations:

$$\exp(\mathbf{A}) = \sum_{k=0}^{\infty} \frac{1}{k!}\mathbf{A}^k \tag{S5}$$

$$[\exp(\mathbf{A})]^{-1} = \exp(-\mathbf{A}) \tag{S6}$$

$$\exp(\mathbf{0}) = \mathbf{I} \tag{S7}$$

$$\log|\exp(\mathbf{A})| = \mathtt{trace}(\mathbf{A}) \tag{S8}$$

(for any square matrix $\mathbf{A}$). Because the vectors $\mathbf{b}_k$ are unit-norm and the trace function is distributive,

$$\log|\mathbf{C}| = \mathtt{trace}(\log \mathbf{C}) = \sum_{jk} \mathtt{trace}(y_j w_{jk} \mathbf{b}_k \mathbf{b}_k^T) = \sum_{jk} y_j w_{jk} \tag{S9}$$

Using the properties above, and plugging the function for the covariance matrix (Eqn. 1) into the

conditional distribution (Eqn. S3) gives

$$\log p(\mathbf{x}|\mathbf{y}) \propto -\frac{1}{2}\sum_{jk} y_j w_{jk} - \frac{1}{2}\mathbf{x}^T \left(\exp\left(-\sum_{jk} y_j w_{jk} \mathbf{b}_k \mathbf{b}_k^T \right) \right) \mathbf{x} \tag{S10}$$

(up to a constant term).

When neural activity is off ($\mathbf{y} = 0$), the covariance matrix is equal to the identity matrix $\mathbf{I}$, corresponding to the canonical distribution of "whitened" images. Non-zero values in neural activity $\mathbf{y}$ scale terms in the sum and thus "warp" the encoded distribution by stretching or contracting along the linear features $\mathbf{b}_k$ (see Fig. 2).

The likelihood function (Eqn. S10 marginalized over all possible values of neural activity $\mathbf{y}$), is maximized to obtain the optimal set of parameters $\mathbf{b}_k$ and w_{jk}. In practice, we evaluate the likelihood at the MAP estimate $\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} p(\mathbf{y}|\mathbf{x})$. The MAP approximation to the integral over $\mathbf{y}$ introduces a degeneracy into learning – the approximate likelihood increases as weights w_{jk} grow without limit while $\hat{\mathbf{y}}$ shrinks – and to address this we fixed the norm of each neuron's weights after an initial period of unconstrained gradient ascent.

Supplementary notes: Additional references

1. de Ruyter van Steveninck, R. and Bialek, W. Real-time performance of a movement sensitive neuron in the blowfly visual system: Coding and information transfer in short spike sequences. *Proc R Soc London Ser B* **234**, 379–414 (1988).

2. Simoncelli, E. P., Pillow, J., Paninski, L., and Schwartz, O. Characterization of neural responses with stochastic stimuli. In The Cognitive Neurosciences, III, Gazzaniga, M., editor, 327–338. MIT Press (2004).

3. Chen, X., Han, F., Poo, M., and Dan, Y. Excitatory and suppressive receptive field subunits in awake monkey primary visual cortex (V1). *Proc. Natl. Acad. Sci. U.S.A.* **104**, 19120–19125, Nov (2007).

4. Rust, N. C., Schwartz, O., Movshon, J. A., and Simoncelli, E. P. Spatiotemporal elements of macaque V1 receptive fields. *Neuron* **46**(6), 945–956 (2005).

5. Adelson, E. H. and Bergen, J. R. Spatiotemporal energy models for the perception of motion. *J Opt Soc Am A* **2**(2), 284–299 (1985).

6. Zhou, Y. X. and Baker, C. L. J. Envelope-responsive neurons in areas 17 and 18 of cat. *J Neurophysiol* **72**(5), 2134–2150 (1994).

7. Mareschal, I. and Baker, C. L. J. Temporal and spatial response to second-order stimuli in cat area 18. *J Neurophysiol* **80**(6), 2811–2823 (1998).

8. David, S. V., Hayden, B. Y., and Gallant, J. L. Spectral receptive field properties explain shape selectivity in area V4. *J Neurophysiology* **96**(6), 3492–505 (2006).

9. Lamme, V. A. The neurophysiology of figure-ground segregation in primary visual cortex. *J Neurosci* **15**(2), 1605–1615 (1995).

10. Lee, T. S., Mumford, D., Romero, R., and Lamme, V. A. The role of the primary visual cortex in higher level vision. *Vision Res* **38**(15-16), 2429–2454 (1998).

11. Nothdurft, H. C., Gallant, J. L., and Van Essen, D. C. Response profiles to texture border patterns in area V1. *Vis Neurosci* **17**(3), 421–436 (2000).

12. Rossi, A. F., Desimone, R., and Ungerleider, L. G. Contextual modulation in primary visual cortex of macaques. *J Neurosci* **21**(5), 1698–1709 (2001).

13. Song, Y. and Baker, C. L. J. Neuronal response to texture- and contrast-defined boundaries in early visual cortex. *Vis Neurosci* **24**(1), 65–77 (2007).

14. Karklin, Y. and Lewicki, M. S. Learning higher-order structures in natural images. *Network: Computation in Neural Systems* **14**, 483–499 (2003).

15. Karklin, Y. and Lewicki, M. S. A hierarchical Bayesian model for learning non-linear statistical regularities in non-stationary natural signals. *Neural Computation* **17**, 397–423 (2005).

16. Olshausen, B. A. and Field, D. J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381**(6583), 607–609 (1996).

17. Bell, A. J. and Sejnowski, T. J. The "independent components" of natural scenes are edge filters. *Vision Res* **37**(23), 3327–3338 (1997).

18. Osindero, S., Welling, M., and Hinton, G. Topographic product models applied to natural scene statistics. *Neural Comput* **18**, 381–414 (2006).

19. Schwartz, O., Sejnowski, T. J., and Dayan, P. Soft mixer assignment in a hierarchical generative model of natural scene statistics. *Neural Computation* **18**(11), 2680–718 (2006).

20. Hinton, G. Learning multiple layers of representation. *Trends in Cognitive Sciences* **11**(10), 428–434, Oct (2007).

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

22. Sahani, M. and Dayan, P. Doubly distributional population codes: simultaneous representation of uncertainty and multiplicity. *Neural Comput* **15**, 2255–2279 (2003).

23. Ma, W. J., Beck, J. M., Latham, P. E., and Pouget, A. Bayesian inference with probabilistic population codes. *Nat Neurosci* **9**(11), 1432–8 (2006).