doi:10.1038/nature07481

Emergence of complex cell properties by learning to generalize in natural scenes

Yan Karklin[1]† & Michael S. Lewicki[1]†

A fundamental function of the visual system is to encode the building blocks of natural scenes—edges, textures and shapes—that subserve visual tasks such as object recognition and scene understanding. Essential to this process is the formation of abstract representations that generalize from specific instances of visual input. A common view holds that neurons in the early visual system signal conjunctions of image features[1,2], but how these produce invariant representations is poorly understood. Here we propose that to generalize over similar images, higher-level visual neurons encode statistical variations that characterize local image regions. We present a model in which neural activity encodes the probability distribution most consistent with a given image. Trained on natural images, the model generalizes by learning a compact set of dictionary elements for image distributions typically encountered in natural scenes. Model neurons show a diverse range of properties observed in cortical cells. These results provide a new functional explanation for nonlinear effects in complex cells[3–6] and offer insight into coding strategies in primary visual cortex (V1) and higher visual areas.

As we scan across a complex natural scene, fixations at multiple locations (for example, on the trunk of a tree or along its edge) produce a coherent percept of the underlying structure (the bark texture or the contour of the edge), even though individual images collected at the retina are inherently highly variable. Figure 1 illustrates the problem our brain solves so effortlessly: perceptually distinct image regions produce response patterns that are highly overlapping and cannot be easily distinguished using low-level, linear representations. What sort of computations are required to achieve generalization across natural stimuli?

Early visual neurons are typically described as linear feature detectors[1,2]. Models developed around this idea can accurately capture the behaviour of neurons from the retina[7] to simple cells in the cortex[8] but, as the examples in Fig. 1 illustrate, neither individual features nor linear transformations can reliably discriminate images of one structure from another. More abstract features are presumably computed in later stages of the visual system, but our knowledge of processing by these neurons is limited. In V1, complex cells respond to an edge over a range of positions[1], but classical models of these cells[9,10] fail to explain a number of nonlinear effects, such as surround suppression and cross-orientation inhibition[3–5]. More importantly, there is no functional explanation for the role of these behaviours in the perception of natural scenes. In higher visual areas such as V2 and V4, neurons are more invariant to image properties such as position and scale[11–13] and might be encoding shape or texture[12,14,15]. For these neurons to generalize effectively, the neural circuitry must generate a representation that is similar across the wide distribution of images of a given type (for example, a texture or contour) yet distinct across the much larger distribution of all other images.

Previous theoretical work has shown that neurons in the primary visual cortex form an efficient code adapted to the statistics of natural images[16,17], but this says nothing about how neurons generalize across

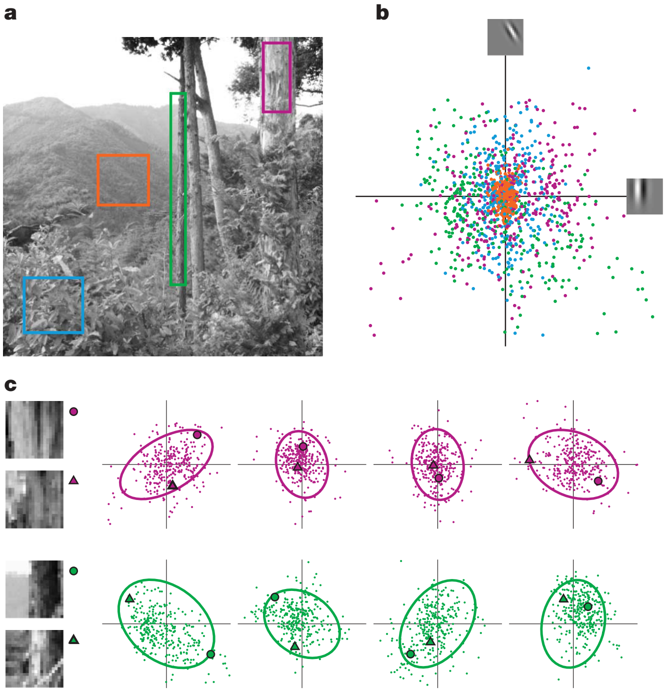

Figure 1 | Statistical patterns distinguish local regions of natural scenes. **a**, A natural scene with four distinct regions outlined (image courtesy of E. Doi). **b**, The scatter plot shows the joint output of a pair of linear feature detectors (oriented Gabor filters) for 20 × 20-image patches sampled from the four regions. The outputs from different regions are highly overlapping, indicating that linear features provide no means to distinguish between the regions. **c**, Each column shows the joint output of a different pair of linear feature detectors sampled from the regions containing the tree bark or the tree edge (the first column corresponds to features in **b**). The correlations in each panel can be described by a Gaussian distribution and its covariance (ellipses). The differences in the distributions between the rows reveal characteristic patterns in correlations, which become even more prominent as projections onto more features are considered. These patterns can be used to generalize within regions while still distinguishing among them. As an example, we highlight two patches in each region, shown by the circle and triangle in each panel. Although the pairs of images are visibly quite different, each image is consistent with the distribution of the local image region. By contrasting the distributions across multiple dimensions, it is possible to infer image type for single patches, even if the patches have similar projections along some image features.

[1]Computer Science Department & Center for the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh, PA 15213, USA. †Present address: Center for Neural Science, New York University, New York, New York, USA (Y.K.); Electrical Engineering and Computer Science Department, Case Western University, Cleveland, Ohio, USA and Wissenschaftskolleg (Institute for Advanced Study) zu Berlin, Germany (M.S.L.).

the intrinsic variability of scene elements. Here we extend the efficient coding approach and propose that an important aspect of visual computation is to represent the myriad statistical distributions that characterize local image regions. Rather than coding the pixel intensities of a patch of texture or edge, neurons in later stages encode the image distribution (that is, the range and pattern of variability of pixel intensities or image features) that is most consistent with the input image. This allows the neural representation to generalize across individual fixations and convey more abstract properties of the image. We demonstrate that a model designed around this computational goal and optimized for natural scenes explains nonlinear properties of complex cells and neurons in higher visual areas, thereby providing a new functional interpretation for these effects.

Fundamentally, generalization is the identification of common characteristics of a class from specific instances. The goal of the proposed model is to learn the statistical distributions that characterize local image regions, such as those in Fig. 1, and identify them from individual image patches. What statistical regularities are relevant for this task? As the examples in Fig. 1 suggest, the distributions of perceptually similar images show consistent patterns in the degree of variation along some dimensions, as well as in the strength of correlations (and anti-correlations) among different feature dimensions. Although these patterns appear subtle when projected onto two dimensions, as in the examples, the full multivariate distribution, consisting of hundreds of dimensions, produces prominent statistical signatures that can be exploited by the visual system.

To determine how the model generalizes, we must specify how it represents distributions of local image regions. A simple way to summarize the patterns of correlations for a given type of image is the covariance matrix of the data. A neural code for this structure could be defined by enumerating the set of observed covariances and assigning one neuron to each pattern, but this approach presents two problems. First, local image classes are not known a priori. Second, given the limited number of neurons in the visual system, it is not feasible to represent all possible image types, let alone the combinatorial number of possible image boundaries. Instead, we propose a distributed code in which the graded activity of the neural population acts to describe a continuum of potential covariance patterns.

This distribution coding model is illustrated schematically in Fig. 2. The model represents the correlations present in local image regions with a multivariate Gaussian distribution that has a fixed mean of zero and a covariance that is a function of the neural activity (see Methods). This simple statistical description affords both the flexibility to capture a continuum of natural image distributions and mathematical simplicity for tractable parameter estimation. The model uses two sets of parameters to describe correlations in image distributions. First, the vectors $\mathbf{b}_k$ (arrows within circles) specify image features along which the encoded distribution can be expanded or contracted relative to the canonical distribution (black circle). These vectors are shared by all neurons in the model (represented by the four grey circles, each of which contains the same set of arrows). Because these vectors do not necessarily line up with the axes of the input dimensions, changes in variation along a vector can correspond to changes in the correlational pattern in many dimensions at once. Neurons in the model (y_j) describe changes along these directions using weights w_{jk}: each has a different set of weights, corresponding to an expansion or contraction along feature $\mathbf{b}_k$. A positive weight (red) means that the neuron responds to a wider range of stimuli along that direction, a negative weight (blue) means it responds to a narrower range, and a weight close to zero (grey) indicates that the neuron is neutral to this direction. The combined activation of all neurons specifies the final shape of the encoded distribution (ellipses). Given a single fixation—one input image—the model computes the neural representation (that is, the image distribution) that provides the most probable explanation of the input. The model is able to generalize over different image regions if the inferred representation is similar across a region (for example, for the pairs of patches in Fig. 2).

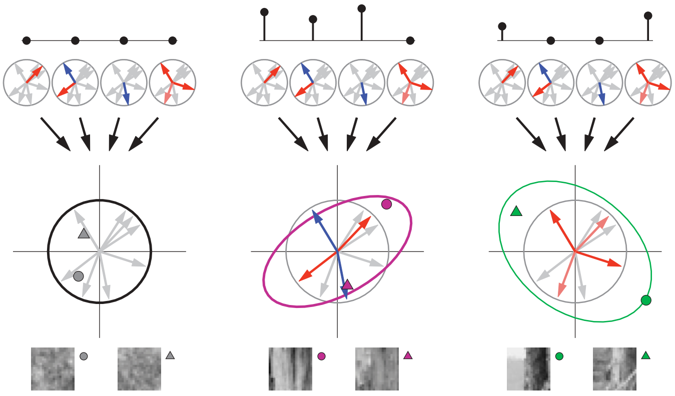

Figure 2 | Distribution coding model. Rather than encoding the precise pixel values of an input image (bottom), the proposed model infers for each image the most likely distribution (ellipses) containing it. Activation patterns for model neurons are shown at the top of each column. Absence of activity corresponds to the lack of image structure (left panel)—that is, a canonical distribution that reflects the statistics over all natural images (black circle). Increased neural activity represents deviations from this canonical distribution and captures statistical patterns in local image regions (middle and right panels, patches and symbols as in Fig. 1). In each panel, the activation pattern is the same for both inputs. See text for further details.

By adapting model parameters $\mathbf{b}_k$ and w_{jk} to the data, we are able to find the most efficient way to use a limited number of neurons to describe the wide range of distributions observed in natural images. It should be noted that, although our goal is to derive a stable representation of all patches within a local region, no assumptions about locality are made (encoding is done independently for each image patch). It is the task of the model to learn a compact representation of all patches and to automatically discover which share the statistical properties of a particular type.

If, as hypothesized, neurons in the visual cortex encode patterns in correlations in local regions and are adapted specifically to the statistics of natural scenes, we expect the representations learned by the model to reflect properties of visual neurons. To this end, we trained the model on patches sampled from a large set of natural images and examined the resulting parameters as well as the response properties of model neurons to natural images.

The vectors $\mathbf{b}_k$ encode the directions of common expansion or contraction in the shape of the image distribution. Drawn as image patches, each is an oriented and localized edge-like feature. The full set tiles the spatial extent of the image patch (Fig. 3a) and spans the range of orientations and spatial frequencies of natural images (not shown). These oriented, band-pass image features are consistent with the optimal images for exciting simple cells in the primary visual cortex[18,19]. Similar representations have been derived previously using linear statistical models that maximize the efficiency of the image codes[16,17]. In the model proposed here, however, these features are not used explicitly to reconstruct the original image, but instead function to modify the encoded distributions (arrows in Fig. 2). Thus, whereas the traditional interpretation of early sensory codes is that they are adapted for faithful reconstruction of the stimulus, our results suggest an additional interpretation: they convey variations in image distributions and allow downstream visual areas to form more abstract representations.

The second set of parameters, the weights w_{jk}, describes the role of each neuron in shaping the encoded image distribution. A set of learned weights for a typical model neuron is shown in Fig. 3b. This neuron exerts the strongest effect on features in the top left of the image patch, increasing the variability (that is, activation) of those oriented at its 'preferred' orientation of 45°, decreasing the variability of those at the orthogonal orientation, as well as those at the preferred orientation but at an offset location. Rather than responding to a few excitatory or suppressive image features, the neuron integrates a large number to describe a pattern of variability underlying a particular image distribution. Although the functional significance of these subunits is to modify the statistical structure of the encoded distribution, they also reflect stimulus features to which this model neuron is most sensitive. It should be noted that a model neuron is activated by all images from this distribution, rather than signalling the presence of one best stimulus. Conversely, stimuli that lie in parts of image space assigned low probability by the neuron inhibit its activity.

To compare the behaviour of the model neuron to that of cells in the visual cortex, we tested its response to stimuli used in classical physiological experiments (sinusoidal gratings). Model parameters were fixed after training on natural images, and neural response computed on a set of patterns centred in the visual area that evoked maximal response. This particular model neuron showed a variety of properties observed in complex cells in V1 and cells in V2, including phase invariance, orientation tuning and complex suppressive effects (Fig. 3c). A large subset of the population exhibits similar properties, whereas others encode more complex patterns that have been observed in higher visual areas V2 and V4 (a detailed analysis of the population and similarities to other experimental data are provided in the Supplementary Information). We emphasize that these results, as well as image features described in Fig. 3a, were obtained with no assumptions about the image structure encoded by visual neurons and without fitting a model to data from physiological experiments. Specifically, we did not restrict the encoded image features to be localized and oriented, nor did we prescribe in advance how the subunits are to be combined in the pattern represented by each neuron.

Finally, we looked at the way in which the model uses the population of neurons to represent images. If the model is able to generalize across the wide variability present in natural images, then image patches that are widely scattered in the original space of linear features should be tightly clustered in the space of the model's representation. This can be illustrated by projecting into two dimensions (as was done with image space in Fig. 1) the model representation of a collection of images (Fig. 4). As hypothesized, by encoding image distributions rather than the precise feature content of each image, model neurons are able to encode perceptually similar images with similar representations and to separate distinct image types.

One limitation of the statistical framework used here is that it does not furnish an explicit feed-forward algorithm for neural encoding. Nevertheless, it is possible to approximate inference in the model by a sequential feed-forward computation: a neuron integrates the squared responses of a large number of image features $\mathbf{b}_k$ and correlates the pattern against its weights w_{jk} (see Supplementary Information for details). This computation can be viewed as a generalization of the standard model of complex cells, in which each complex cell takes as input the squared output of two simple cells[9,10,20,21]. In contrast, model neurons can receive many inputs, and the linear features themselves are learned. We find that the optimal number of input features varies greatly, and the features are integrated in a variety of ways. These predictions are consistent with recent analyses of functional subfields in V1 complex cells[6,22]. In addition, some model neurons integrate more complex spatial patterns (see Supplementary Information), which predicts a neural response to a richer variety of images than has been tested

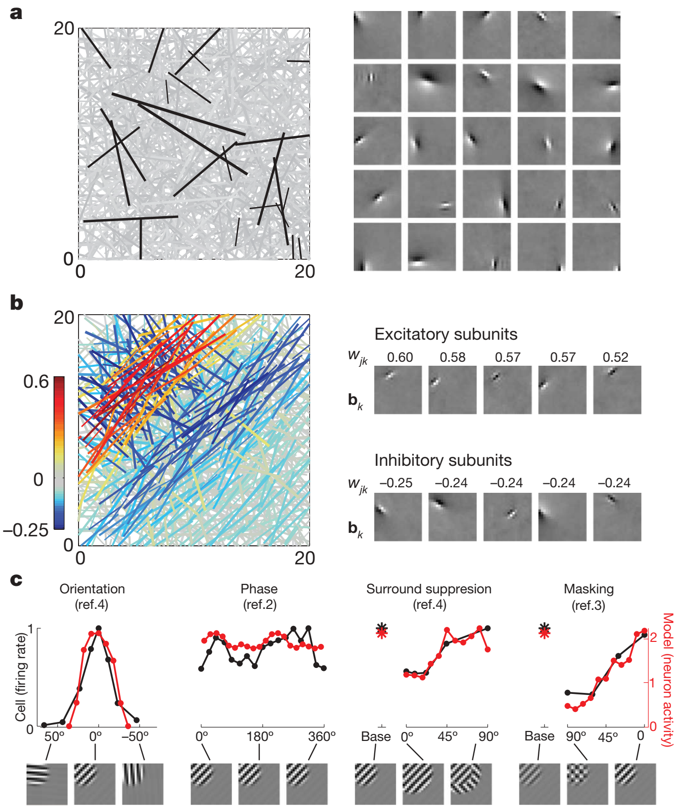

Figure 3 | Model neurons exhibit properties of cortical visual neurons. **a**, When adapted to natural images, the vectors $\mathbf{b}_k$ are oriented, localized in space, and span the spatial extent of the 20 × 20-pixel image patch. Each line reflects the orientation, spatial position within the image patch, and scale (length of line) of one of the image features. Twenty-five representative features (from a total of 1,000) are drawn in black, and shown in image form on the right. **b**, Weights of one typical model neuron to the features $\mathbf{b}_k$. As in **a**, each feature is represented by a line, and the colour of the line indicates the sign and magnitude of the weight to the feature (see colour bar). Positive weights indicate increased variability in the corresponding feature; negative weights indicate decreased variability; features to which the neuron is insensitive are shaded grey. Image features ($\mathbf{b}_k$) corresponding to the five most positive and the five most negative weights are shown in the right panel; the corresponding weights are above each feature. These act as excitatory and inhibitory subunits for this neuron. **c**, When presented with sinusoidal gratings, this model neuron replicates common aspects of the neural response in complex cells in cortical area V1. It is highly tuned to the grating's orientation, but insensitive to its phase. Adding a grating into the surrounding region suppresses the response (third plot, 0°) relative to baseline response to a single grating (asterisk), but this suppression is tuned to the orientation of the surround and is weakest when the surround is orthogonal to the preferred orientation (90°). Masking with a superimposed orthogonal grating suppresses the response (fourth plot, 90°), but this suppression is also orientation-dependent. All model neuron responses are plotted on the same scale (red axis); cell firing rates in each plot were normalized to a maximum value of 1; preferred orientation was shifted to 0° for the model neuron and the cell in all plots.

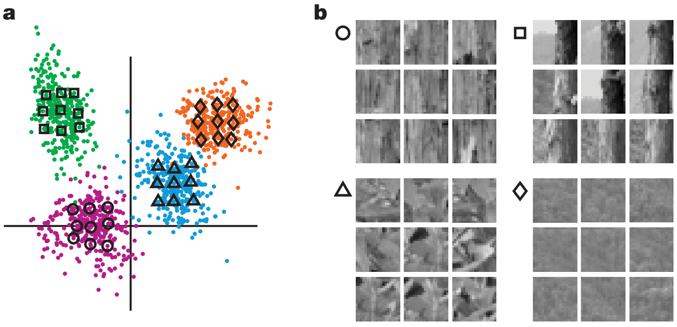

Figure 4 | Generalization across natural variability. **a**, In contrast to linear projections (compare to Fig. 1b), a two-dimensional projection of the model's representation (the activity of 150 model neurons) reveals well-separated clusters. **b**, Each 3 × 3-image group corresponds to the array of symbols in **a**. Despite the variability in the appearance of edges and textures, the model's representation of natural images generalizes within each region while still distinguishing among them.

physiologically. Experiments that incorporate such stimuli will provide an important validation of the proposed model.

The nonlinear effects shown by neurons in the model (Fig. 3c) have been previously incorporated into models of complex cells[5,8,20,21]. Much of this work has focused on fitting mathematical models to neural data[5,8,20,23] and does not provide a functional explanation of the observed neural properties. Other models have been motivated by specific computational goals, such as statistical independence[24,25], stability of representation over time[26,27], or position or scale invariance[28]. However, these models do not explicitly address the problem of generalization, which here is performed by inferring the statistical distribution that is most likely to explain the input image. An important advantage of our approach is that, rather than assuming invariance (or sensitivity) to limited stimulus parameters such as position or orientation, the model learns a much more general set of features that are determined by the statistical structures in natural images. If higher-level visual neurons are generalizing according to these statistics, they should have invariance along specific stimulus dimensions, and their responses to natural images should reflect common statistical structure in local image regions. Thus, the model provides a quantitative way to explore neural responses to complex stimuli characterized by their statistical regularities.

METHODS SUMMARY

The model describes individual image patches $\mathbf{x}$ with multivariate Gaussian probability distributions:

$$P(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\boldsymbol{\mu}, \mathbf{C}) \qquad (1)$$

with mean $\boldsymbol{\mu} = 0$ and with covariance a function of the neural encoding vector $\mathbf{C} = f(\mathbf{y})$. The logarithm of the covariance matrix is given by the combination of outer products of feature vectors $\mathbf{b}_k$, weighted by neural activities y_j through weights w_{jk}:

$$\log \mathbf{C} = \sum_{jk} y_j w_{jk} \mathbf{b}_k \mathbf{b}_k^T \qquad (2)$$

Because a different covariance can be inferred for each image, the distribution over the entire ensemble of images is highly non-Gaussian. (This model is a generalized version of the hierarchical model described previously[29], which captured patterns among the variances, but not the correlations, of linear features.)

We trained the model on a large set of 20 × 20 image patches, sampled randomly from greyscale photographs of outdoor scenes[19]. The number of neurons was set to 150 and the number of linear features $\mathbf{b}_k$ to 1,000. The 'response' of model neurons was computed as the most probable neural representation given the input image by maximizing the posterior probability $P(\mathbf{y}|\mathbf{x}, \{\mathbf{b}_k, w_{jk}\})$. Model parameters were initialized to small random values and optimized by maximizing the likelihood of the data under the model $P(\mathbf{x}|\{\mathbf{b}_k, w_{jk}\})$ using standard gradient ascent.

For the 'physiological' analysis of Fig. 3c, we first identified the location, orientation, and spatial extent and frequency of a windowed sinusoidal grating that best activated the model neuron (one that yielded the most positive value of y_j). We then varied each tested parameter and computed the model's representation of the stimulus (the vector of responses of model neurons).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 4 May; accepted 26 September 2008.
Published online 19 November 2008.

1. Hubel, D. H. & Wiesel, T. N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol. (Lond.)* **160,** 106–154 (1962).
2. Movshon, J. A., Thompson, I. D. & Tolhurst, D. J. Spatial summation in the receptive fields of simple cells in the cat's striate cortex. *J. Physiol. (Lond.)* **283,** 53–77 (1978).
3. Bonds, A. B. Role of inhibition in the specification of orientation selectivity of cells in the cat striate cortex. *Vis. Neurosci.* **2,** 41–55 (1989).
4. Jones, H. E., Wang, W. & Sillito, A. M. Spatial organization and magnitude of orientation contrast interactions in primate V1. *J. Neurophysiol.* **88,** 2796–2808 (2002).
5. Cavanaugh, J. R., Bair, W. & Movshon, J. A. Nature and interaction of signals from the receptive field center and surround in macaque V1 neurons. *J. Neurophysiol.* **88,** 2530–2546 (2002).
6. Chen, X., Han, F., Poo, M. & Dan, Y. Excitatory and suppressive receptive field subunits in awake monkey primary visual cortex (V1). *Proc. Natl Acad. Sci. USA* **104,** 19120–19125 (2007).
7. Chichilnisky, E. J. A simple white noise analysis of neuronal light responses. *Network: Comp. Neural Syst.* **12,** 199–213 (2001).
8. Carandini, M., Heeger, D. J. & Movshon, J. A. Linearity and normalization in simple cells of the macaque primary visual cortex. *J. Neurosci.* **17,** 8621–8644 (1997).
9. Movshon, J. A., Thompson, I. D. & Tolhurst, D. J. Receptive field organization of complex cells in the cat's striate cortex. *J. Physiol. (Lond.)* **283,** 79–99 (1978).
10. Adelson, E. H. & Bergen, J. R. Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am. A* **2,** 284–299 (1985).
11. Kobatake, E. & Tanaka, K. Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *J. Neurophysiol.* **71,** 856–867 (1994).
12. Gallant, J. L., Connor, C. E., Rakshit, S., Lewis, J. W. & Van Essen, D. C. Neural responses to polar, hyperbolic, and Cartesian gratings in area V4 of the macaque monkey. *J. Neurophysiol.* **76,** 2718–2739 (1996).
13. Connor, C. E., Brincat, S. L. & Pasupathy, A. Transformation of shape information in the ventral pathway. *Curr. Opin. Neurobiol.* **17,** 140–147 (2007).
14. Hegdé, J. & Van Essen, D. C. Selectivity for complex shapes in primate visual area V2. *J. Neurosci.* **20,** RC61:1–6 (2000).
15. Pasupathy, A. & Connor, C. E. Shape representation in area V4: position-specific tuning for boundary conformation. *J. Neurophysiol.* **86,** 2505–2519 (2001).
16. Olshausen, B. A. & Field, D. J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381,** 607–609 (1996).
17. Bell, A. J. & Sejnowski, T. J. The ''independent components'' of natural scenes are edge filters. *Vision Res.* **37,** 3327–3338 (1997).
18. Jones, J. P. & Palmer, L. A. The two-dimensional spatial structure of simple receptive fields in cat striate cortex. *J. Neurophysiol.* **58,** 1187–1211 (1987).
19. van Hateren, J. H. & van der Schaaf, A. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc. R. Soc. Lond. B* **265,** 359–366 (1998).
20. Heeger, D. J. Normalization of cell responses in cat striate cortex. *Vis. Neurosci.* **9,** 181–197 (1992).
21. Heeger, D. J., Simoncelli, E. P. & Movshon, J. A. Computational models of cortical visual processing. *Proc. Natl Acad. Sci. USA* **93,** 623–627 (1996).
22. Rust, N. C., Schwartz, O., Movshon, J. A. & Simoncelli, E. P. Spatiotemporal elements of macaque V1 receptive fields. *Neuron* **46,** 945–956 (2005).
23. Cadieu, C. *et al.* A model of V4 shape selectivity and invariance. *J. Neurophysiol.* **98,** 1733–1750 (2007).
24. Schwartz, O. & Simoncelli, E. P. Natural signal statistics and sensory gain control. *Nature Neurosci.* **4,** 819–825 (2001).
25. Hyvärinen, A. & Hoyer, P. A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images. *Vision Res.* **41,** 2413–2423 (2001).
26. Berkes, P. & Wiskott, L. Slow feature analysis yields a rich repertoire of complex cell properties. *J. Vis.* **5,** 579–602 (2005).
27. Hurri, J. & Hyvärinen, A. Simple-cell-like receptive fields maximize temporal coherence in natural video. *Neural Comput.* **15,** 663–691 (2003).
28. Riesenhuber, M. & Poggio, T. Hierarchical models of object recognition in cortex. *Nature Neurosci.* **2,** 1019–1025 (1999).
29. Karklin, Y. & Lewicki, M. S. A hierarchical Bayesian model for learning non-linear statistical regularities in non-stationary natural signals. *Neural Comput.* **17,** 397–423 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported by the Department of Energy through the Computational Science Graduate Fellowship (to Y.K.), the National Science Foundation Grant under grant numbers 0413152 and 0705677 (to M.S.L.) and the Office of Naval Research under the Multidisciplinary University Research Initiative N000140710747.

Author Contributions Y.K. and M.S.L. developed the model, analysed the results and wrote the paper; Y.K. ran the simulations.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to Y.K. (yan.karklin@nyu.edu) or M.S.L. (michael.lewicki@case.edu).

METHODS

Data. We used 110 greyscale images of outdoor scenes as training data[19]. Pixel intensities were log-transformed (corresponding roughly to the transformation at the retinal cone cells[30]), and the images were low-pass filtered to remove corner frequency sampling artefacts. We randomly extracted overlapping 20×20-image patches from the entire data set. The mean luminance value was subtracted from each patch (which sped up model training but had no significant influence on the results). We 'whitened' all image patches to remove global correlations and to normalize the variance; this allowed the model to encode only the deviations of each image distribution from the global statistics (the canonical distribution). For visualization of image features, the results were projected back into the original image space. All stimuli in the physiological analysis of Fig. 3c were preprocessed in the same way as the natural images used in training.

Model parameter estimation. We estimated the optimal model parameters $\theta = \{\mathbf{b}_k, w_{jk}\}$ by maximizing the likelihood of the data under the model

$$P(\mathbf{x}|\theta) = \int P(\mathbf{x}|\mathbf{y}, \theta)P(\mathbf{y})d\mathbf{y} \qquad (3)$$

The conditional distribution $P(\mathbf{x}|\mathbf{y}, \theta)$ is a multivariate Gaussian that captures correlations in local image regions (equation (1)). Neural activities were assumed to be sparse[31] and independent, and were modelled with a Laplacian (symmetric exponential) distribution, $P(\mathbf{y}) = \Pi P(y_j) \propto \Pi e^{-|y_j|}$. The integral over all possible neural states in equation (3) is intractable and was replaced by a single evaluation at the maximum a posteriori value $\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} P(\mathbf{y}|\mathbf{x}, \theta)$. Although this approximation ignores the volume around the maximum, it is one standard approach to tackling this problem.

We assumed the training patches were sampled independently and that the likelihood for the data ensemble was a product of terms for individual images (equation (3)). In practice, we maximized the log-likelihood using gradient ascent on batches of 100 image patches. Repeated training runs produced convergence to similar parameter values.

Model responses to grating stimuli. The orientation tuning of model neurons in Fig. 3c was measured using 20×20 patches of sinusoidal gratings at different positions, orientations, spatial frequencies and phases. We first eliminated neurons that were 'unresponsive' to gratings, that is, those whose maximal response did not reach 2 standard deviations of the population response to gratings. This was necessary to discount small random activation of neurons tuned for other types of image structures. For each neuron we found the grating with the maximal response and measured modulation in response to varying orientation, phase, or the addition of masks in the receptive field or the surround. Because neural activity could be positive or negative, the full amplitude of modulation was considered as twice the maximum absolute value of the response.

A neuron was considered to be orientation-tuned if its response was modulated by more than 50% over the range of stimulus orientations, and to be phase invariant if the response varied less than 50% over phase-shifted gratings. Cross-orientation inhibition and surround suppression corresponded to greater than 25% decrease in neural response. Bandwidth of orientation tuning was computed as the width at $1/\sqrt{2}$ of the full amplitude of the response modulation.

The projection of neural activity in Fig. 4 was computed using linear discriminant analysis, a technique that finds the linear projections that best separate different classes of data. Applied to the raw pixel data or to the outputs of linear features (data shown in Fig. 1), this method failed to separate the clusters.

30. van Hateren, J. H. Processing of natural time series of intensities by the visual system of the blowfly. *Vision Res.* **37,** 3407–3416 (1997).

31. Olshausen, B. A. & Field, D. J. Sparse coding of sensory inputs. *Curr. Opin. Neurobiol.* **14,** 481–487 (2004).

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 suppresive 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

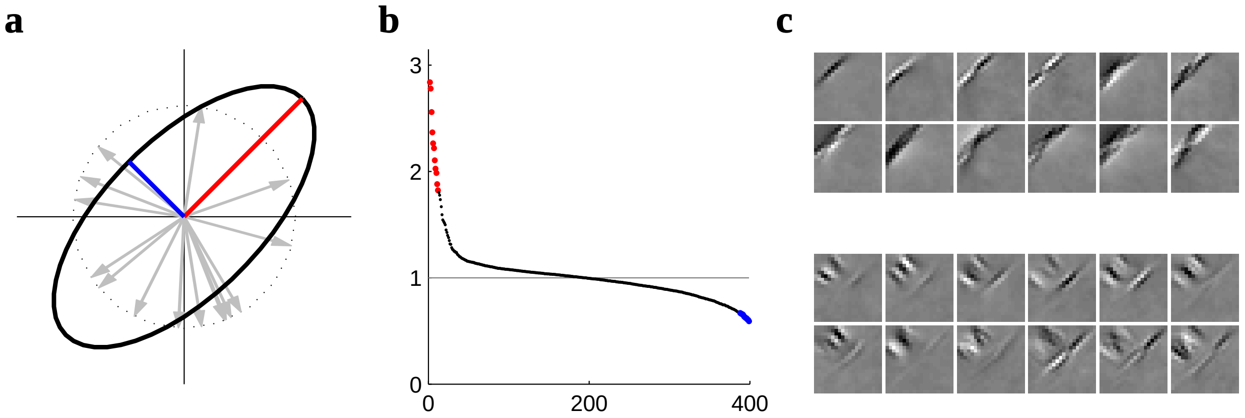

Figure S1: **a**. A schematic of one model neuron's effect on the encoded distribution. The neuron uses the underlying image features (gray arrows) to transform the canonical distribution (dotted circle) into a different distribution (black ellipse). The effect of the neuron on the distribution is given 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-

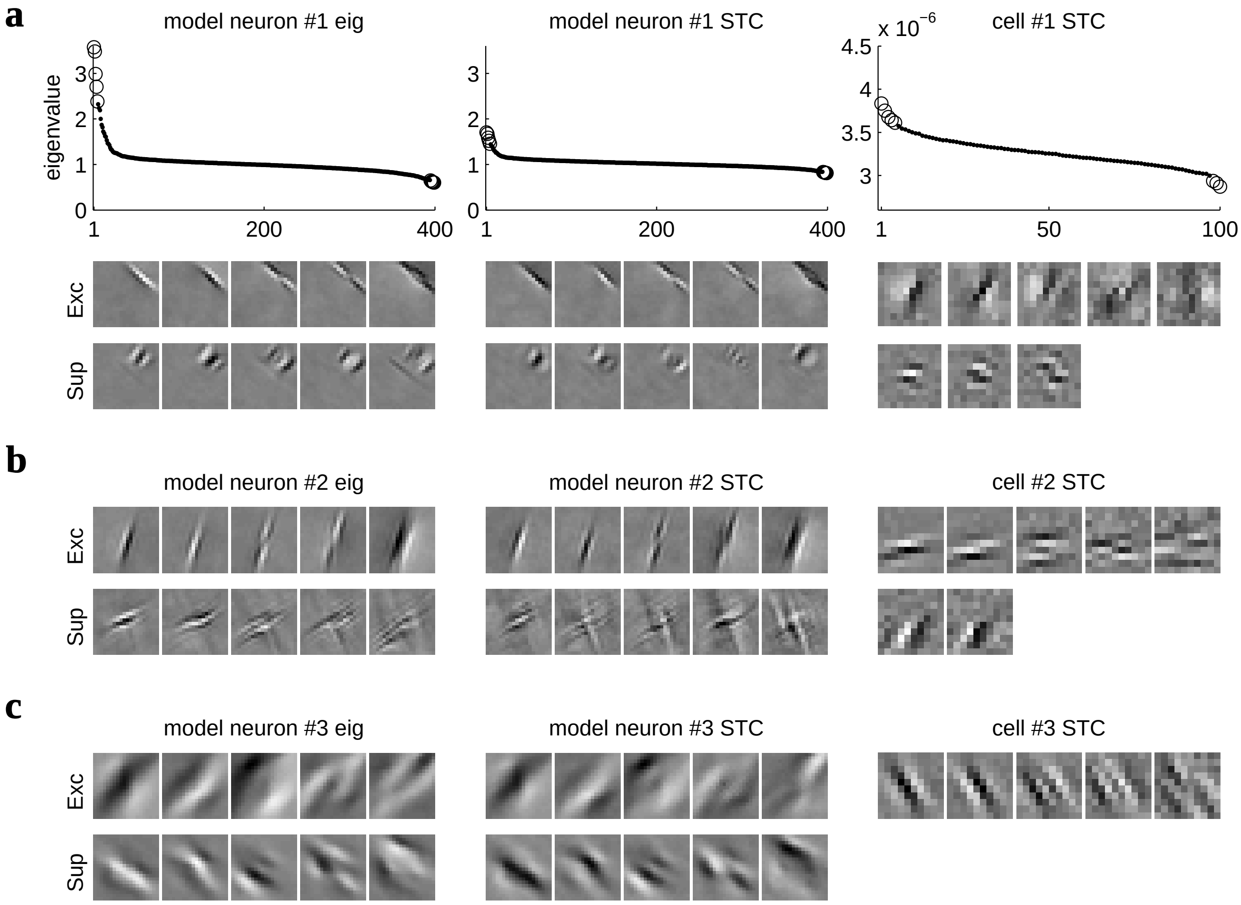

Figure S2: **Spike-triggered characterization of model neurons reveals functional subunits similar to those found in V1 complex cells.** **a**. Eigenvalues and eigenvectors for one model neuron, as revealed by the eigenvector analysis of model parameters (first column) and STC of model responses to white 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

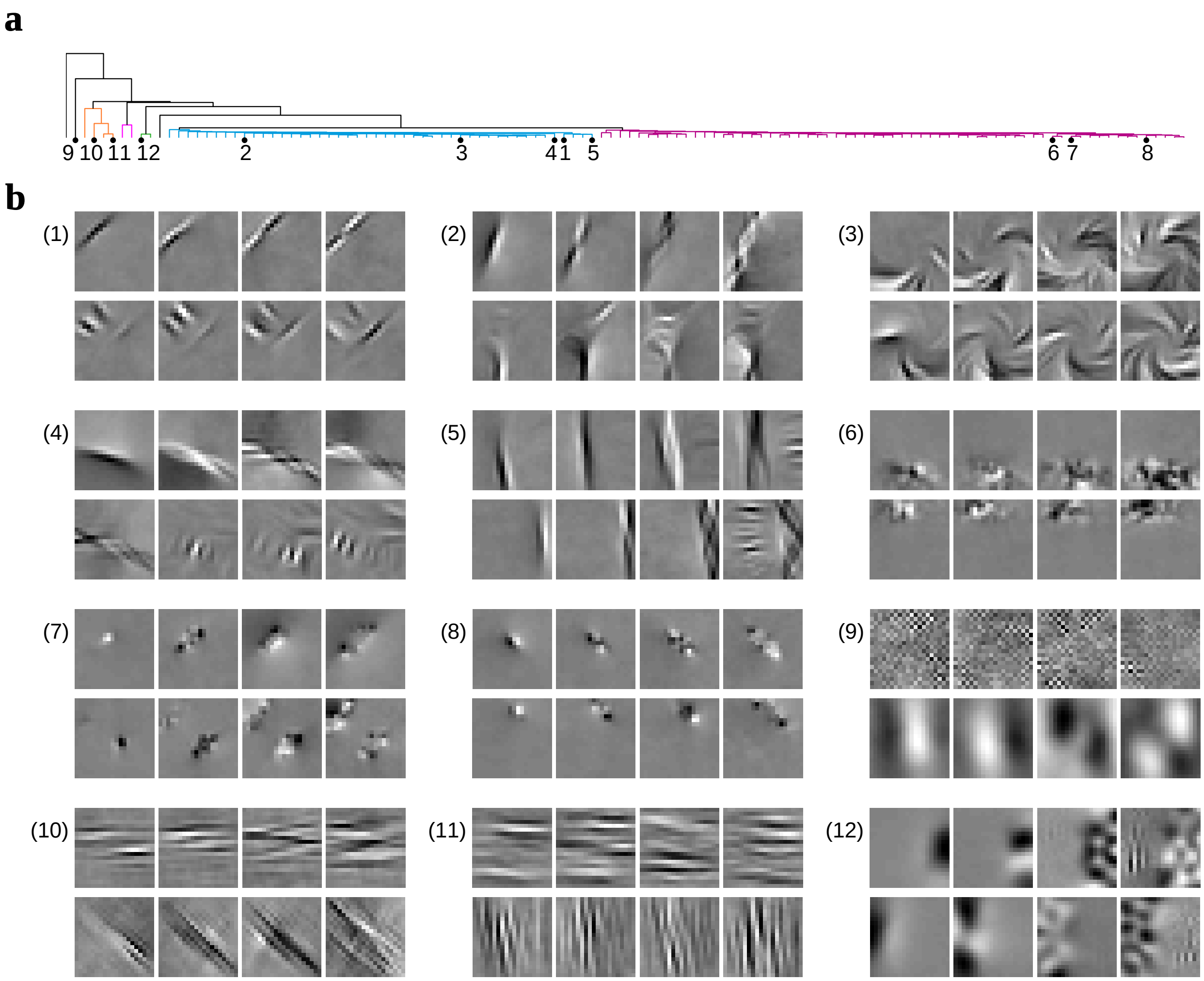

Figure S3: **Image distributions are encoded by a diverse population of neurons.** **a**. 120 most active (out of a total of 150) neurons were hierarchically clustered according to the different aspects of image structure they encode (see text for details). The clustering reveals two large categories 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:

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

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).