Soft Mixer Assignment in a Hierarchical Generative Model of Natural Scene Statistics

Odelia Schwartz
odelia@salk.edu
Howard Hughes Medical Institute, Computational Neurobiology Lab, Salk Institute for Biological Studies, La Jolla, CA 92037, U.S.A.

Terrence J. Sejnowski
terry@salk.edu
Howard Hughes Medical Institute, Computational Neurobiology Lab, Salk Institute for Biological Studies, La Jolla, CA 92037, and Department of Biology, University of California at San Diego, La Jolla, CA 92093, U.S.A.

Peter Dayan
dayan@gatsby.ucl.ac.uk
Gatsby Computational Neuroscience Unit, University College, London WC1N 3AR, U.K.

Gaussian scale mixture models offer a top-down description of signal generation that captures key bottom-up statistical characteristics of filter responses to images. However, the pattern of dependence among the filters for this class of models is prespecified. We propose a novel extension to the gaussian scale mixture model that learns the pattern of dependence from observed inputs and thereby induces a hierarchical representation of these inputs. Specifically, we propose that inputs are generated by gaussian variables (modeling local filter structure), multiplied by a mixer variable that is assigned probabilistically to each input from a set of possible mixers. We demonstrate inference of both components of the generative model, for synthesized data and for different classes of natural images, such as a generic ensemble and faces. For natural images, the mixer variable assignments show invariances resembling those of complex cells in visual cortex; the statistics of the gaussian components of the model are in accord with the outputs of divisive normalization models. We also show how our model helps interrelate a wide range of models of image statistics and cortical processing.

1 Introduction

The analysis of the statistical properties of natural signals such as photographic images and sounds has exerted an important influence over both

Neural Computation **18, 2680–2718** (2006)

sensory systems neuroscience and signal processing. From the earliest days of the electrophysiological investigation of the neural processing of visual input, it has been hypothesized that neurons in early visual areas decompose natural images in a way that is sensitive to aspects of their probabilistic structure (Barlow, 1961; Attneave, 1954; Simoncelli & Olshausen, 2001). The same statistics lie at the heart of effective and efficient methods of image processing and coding.

There are two main approaches to the study of the statistics of natural signals. Bottom-up methods start by studying the empirical statistical regularities of various low-dimensional linear or nonlinear projections of the signals. These methods see cortical neurons in terms of choosing and manipulating projections, to optimize probabilistic and information-theoretic metrics (Shannon, 1948; Shannon & Weaver, 1949), such as sparsity (Field, 1987), and efficient coding including statistical independence (Barlow, 1961; Attneave, 1954; Li & Atick, 1994; Nadal & Parga, 1997). In contrast, top-down methods (Neisser, 1967; Hinton & Ghahramani, 1997) are based on probabilistic characterizations of the processes by which the signals are generated and see cortical neurons as a form of coordinate system parameterizing the statistical manifold of the signals.

There has been substantial recent progress in bottom-up statistics. In particular, a wealth of work has examined the statistical properties of the activation of linear filters convolved with images. The linear filters are typically chosen to qualitatively match retinal or cortical receptive fields. For example, primary visual cortex receptive fields (e.g., simple cells) are tuned to a localized spatial region, orientation, and spatial frequency (Hubel & Wiesel, 1962). These receptive fields are also closely related to multiscale wavelet decompositions, which have gained wide acceptance in the computational vision community. For typical natural images, empirical observations of a single linear filter activation reveal a highly kurtotic (e.g., sparse) distribution (Field, 1987). Groups of linear filters (coordinated across parameters such as orientation, frequency, phase, or spatial position) exhibit a striking form of statistical dependency (Wegmann & Zetzsche, 1990; Zetzsche, Wegmann, & Barth, 1993; Simoncelli, 1997), which can be characterized in terms of the variance (Simoncelli, 1997; Buccigrossi & Simoncelli, 1999; Schwartz & Simoncelli, 2001). The importance of variance statistics had been suggested earlier in pixel space (Lee, 1980) and has been addressed in other domains such as speech (Brehm & Stammler, 1987) and even finance (Bollerslev, Engle, & Nelson, 1994).

There has also been substantial recent progress in top-down methods (Rao, Olshausen, & Lewicki, 2002), especially in understanding the tight relationship between bottom-up and top-down ideas. In particular, it has been shown that optimizing a linear filter set for statistical properties such as sparseness or marginal independence (Olshausen & Field, 1996; Bell & Sejnowski, 1997; van Hateren & van der Schaaf, 1998) in the light of the statistics of natural images can be viewed as a way of fitting an exact or approximate top-down generative model (Olshausen & Field, 1996). These methods all lead to optimal filters that are qualitatively matched to simple cells. The bottom-up variance coordination among the filters has also found a resonance in top-down models (Wainwright & Simoncelli, 2000; Wainwright, Simoncelli, & Willsky, 2001; Hyvärinen & Hoyer, 2000a; Romberg, Choi, & Baraniuk, 1999, 2001; Karklin & Lewicki, 2003a, 2005). Various generative models have built hierarchies on top of simple cell receptive fields, leading to nonlinear cortical properties such as the phase invariance exhibited by complex cells together with other rich invariances.

This article focuses on a hierarchical, nonlinear generative modeling approach to understanding filter coordination and its tight relation to bottom-up statistics. We build on two substantial directions in the literature, whose close relationship is only slowly being fully understood.

One set of ideas started in the field of independent component analysis (ICA), adding to the standard single, linear layer of filters a second layer that determines the variance of the first-layer activations (Hyvärinen & Hoyer, 2000a, 2000b; Hoyer & Hyvärinen, 2002; Karklin & Lewicki, 2003a, 2005; Park & Lee, 2005). In particular, Karklin and Lewicki (2003a, 2003b, 2005) suggested a model in which the variance of each unit in the first layer arises from an additive combination of a set of variance basis function units in the second layer. The method we propose can be seen as a version of this with competitive rather than cooperative combination of the second-layer units.

The other set of ideas originates with the gaussian scale mixture model (GSM) (Andrews & Mallows, 1974; Wainwright & Simoncelli, 2000; Wainwright et al., 2001),[1] which has strong visibility in the image processing literature (Strela, Portilla, & Simoncelli, 2000; Portilla, Strela, Wainwright, & Simoncelli, 2001, 2003; Portilla & Simoncelli, 2003). GSM generative models offer a simple way of parameterizing the statistical variance dependence of the first-layer filter activations in a way that captures some of the key bottom-up statistical properties of images. However, although GSMs parameterize the dependence of linear filters, they do not by themselves specify the pattern of dependence among the filters. This is the key hurdle in their application as a top-down basis for bottom-up, hierarchical learning models. In these terms, we propose an extension to the GSM model that learns the pattern of dependencies among linear filters, thereby learning a hierarchical representation.

In the next section, we discuss bottom-up statistical properties of images. We describe and motivate the use of gaussian scale mixture models and then pose the question of learning a hierarchical representation in this framework. This lays the groundwork for the rest of the article, in which we develop the model and hierarchical learning more formally and demonstrate results on both synthetic data and natural image ensembles.

[1] Another class of models, which has recently been related both to ICA and the GSM, is the energy-based product of Student-t models (Osindero et al., 2006).

An earlier version of part of this work appeared in Schwartz, Sejnowski, & Dayan (2005).

2 Bottom-Up and Top-Down Statistics of Images

At the heart of both bottom-up and top-down methods are the individual and joint statistics of the responses of the set of linear Gabor-like filters that characterize simple-cell receptive fields in primary visual cortex.

The distribution of the activation of a single linear filter when convolved with an image is highly kurtotic. That is, the response of the filter is often approximately zero, but occasionally the filter responds strongly to particular structures in the image (Field, 1987).

The joint statistics of two related linear filters convolved with the same image exhibit a striking form of statistical dependence: when one of the filters responds strongly to a prominent aspect in the image, the other filter may also respond strongly (say, if two spatially displaced vertical filters are responding to an elongated vertical edge in the image). This is also known as a self-reinforcing characteristic of images (e.g., Turiel, Mato, Parga, & Nadal, 1998). The strength of this dependence is determined by the featural similarity of the linear filters in terms of relative location, orientation, spatial scale, phase, and so forth. The coordination is reflected in the joint conditional distribution having the shape of a bowtie and thus following a variance dependency (Buccigrossi & Simoncelli, 1999; Schwartz & Simoncelli, 2001), or by examining the marginal versus the joint distributions (Zetzsche et al., 1993; Zetzsche & Nuding, 2005). Huang and Mumford (1999) analyzed joint contour plots for a large image database and modeled the joint dependencies as a generalized 2D gaussian. The dependencies can be seen in the responses of various types of linear filters, including predefined wavelets and filters designed to be maximally sparse or independent. These are also present even when the filter responses are linearly decorrelated.

Another view on this self-reinforcing characteristics comes (Wainwright & Simoncelli, 2000; Wainwright et al., 2001) from the top-down GSM model, which was originally described by Andrews and Mallows (1974) over 30 years ago. The model consists of two components: a multidimensional gaussian $\mathbf{g}$, multiplied by a positive scalar random variable v. The second component v effectively "scales" the gaussian component $\mathbf{g}$, forming a "mixture," $\mathbf{l}$, according to the equation:

$$\mathbf{l} = v\mathbf{g} \tag{2.1}$$

with density

$$p[\mathbf{l}] = \int \frac{1}{(2\pi)^{m/2}|v^2\Sigma|^{1/2}} \exp\left(-\frac{\mathbf{l}^t\Sigma^{-1}\mathbf{l}}{2v^2}\right) p[v]dv, \tag{2.2}$$

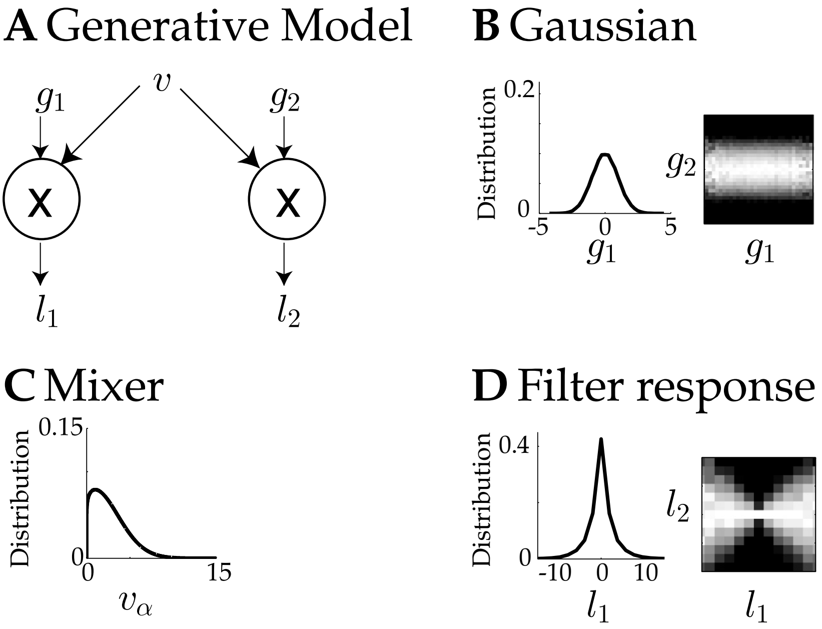


Figure 1: (A) Generative model for a two-dimensional GSM. Each filter response, l_1 and l_2, is generated by multiplying (circle with X symbol) its gaussian variable, g_1 and g_2, by a common mixer variable v. (B) Marginal and joint conditional statistics (bowties) of the gaussian components of the GSM. For the joint conditional statistics, intensity is proportional to the bin counts, except that each column is independently rescaled to fill the range of intensities. (C) Marginal statistics of the mixer component of the GSM. The mixer is by definition positive and is chosen here from a Rayleigh distribution with parameter $a = .1$ (see equation 3.1), but exact distribution of mixer is not crucial for obtaining statistical properties of filter responses shown in D. (D) Marginal and joint conditional statistics (bowties) of generated filter responses.

where m is the number of filters, Σ is the covariance matrix, and mixer v is distributed according to a prior distribution $p[v]$.[2]

In its application to natural images (Wainwright & Simoncelli, 2000), we typically think of each l_i as modeling the response of a single linear filter when applied to a particular image patch. We will also use the same analogy in describing synthetic data. We refer to the scalar variable v as a

[2] In other work, the mixture has also been defined as $\mathbf{l} = \sqrt{v}\mathbf{g}$, resulting in slightly different notation.

mixer variable to avoid confusion with the scales of a wavelet.[3] Figure 1A illustrates a simple two-dimensional GSM generative model, in which l_1 and l_2 are generated with a common mixer variable v. Figures 1B and 1C show the marginal and joint conditional gaussian statistics of the gaussian and mixer variables for data synthesized from this model.

The GSM model provides the top-down account of the two bottom-up characteristics of natural scene statistics described earlier: the highly kurtotic marginal statistics of a single linear filter and the joint conditional statistics of two linear filters that share a common mixer variable (Wainwright & Simoncelli, 2000; Wainwright et al., 2001). Figure 1D shows the marginal and joint conditional statistics of two filter responses l_1 and l_2 based on the synthetic data of Figures 1B and 1C.

The GSM model bears a close relationship with bottom-up approaches of image statistics and cortical representation. First, models of sparse coding and cortical receptive field representation typically utilize the leptokurtotic properties of the marginal filter response, which arise naturally in a generative GSM model (see Figure 1D, left). Second, GSMs offer an account of filter coordination, as in, for instance, the bubbles framework of Hyvärinen (Hyvärinen, Hurri, & Vayrynen, 2003). Coordination arises in the GSM model when filter responses share a common mixer (see Figure 1D, right). Third, some bottom-up frameworks directly consider versions of the two GSM components. For instance, models of image statistics and cortical gain control (Schwartz & Simoncelli, 2001) result in a divisively normalized output component that has characteristics resembling that of the gaussian component of the GSM in terms of both the marginal and joint statistics (see Figure 1B and Wainwright & Simoncelli, 2000). Further, Ruderman and Bialek (1994) postulate that the observed pixels in an image (note, *not* the response of linear filters convolved with an image) can be decomposed into a product of a local standard deviation and a roughly gaussian component. In sum, the GSM model offers an attractive way of unifying a number of influential statistical approaches.

In the original formulation of a GSM, there is one mixer for a single collection of gaussian variables, and their bowtie statistical dependence is therefore homogeneous. However, the responses of a whole range of linear filters to image patches are characterized by heterogeneity in their degrees of statistical dependence. Wainwright and Simoncelli (2000) considered a prespecified tree-based hierarchical arrangement (and indeed generated the mixer variables in a manner that depended on the tree). However, for a diverse range of linear filters and a variety of different classes of scenes, it is necessary to learn the hierachical arrangement from examples. Moreover, because different objects induce different dependencies, different

[3] Note that in some literature, the scalar variable has also been called a multiplier variable.

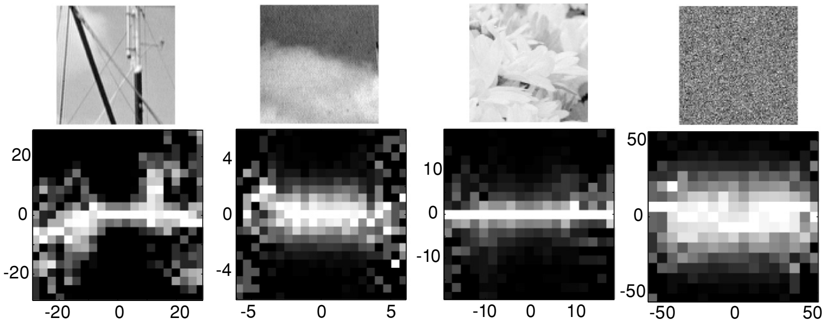


Figure 2: Joint conditional statistics for different image patches, including white noise. Statistics are gathered for a given pair of vertical filters that are spatially nonoverlapping. Image patches are 100 by 100 pixels. Intensity is proportional to the bin counts, except that each column is independently rescaled to fill the range of intensities.

arrangements may be appropriate for different image patches. For example, for a given pair of filters, the strength of the joint conditional dependency can vary for different image patches (see Figure 2). This suggests that on a patch-by-patch basis, different mixers should be associated with different filters. Karklin and Lewicki (2003a) suggested what can be seen as one way of doing this: generating the (logarithm of the) mixer value for each filter as a linear combination of the values of a small number of underlying mixer components.

Here, we consider the problem in terms of multiple mixer variables $\mathbf{v} = (v_\alpha, v_\beta \ldots)$, with the linear filters being clustered into groups that share a single mixer. As illustrated in Figure 3, this induces an assignment problem of marrying linear filter responses l_i and mixers v_j, which is the main focus of this article. Inducing the assignment is exactly the process of inducing a level of a hierarchy in the statistical model. Although the proposed model is more complex than the original GSM, in fact we show that inference is straightforward using standard tools of expectation maximization (Dempster, Laird, & Rubin, 1977) and Markov chain Monte Carlo sampling. Closely related assignment problems have been posed and solved using similar techniques, in a different class of image model known as dynamical tree modeling (Williams & Adams, 1999; Adams & Williams, 2003) and in credibility networks (Hinton, Ghahramani, & Teh, 1999).

In this article, we approach the question of hierarchy in the GSM model. In section 3, we consider estimating the gaussian and mixer variables of a GSM model from synthetic and natural data. We show how inference fails in the absence of correct knowledge about the assignment associations between gaussian and mixer variables that generated the data. For this

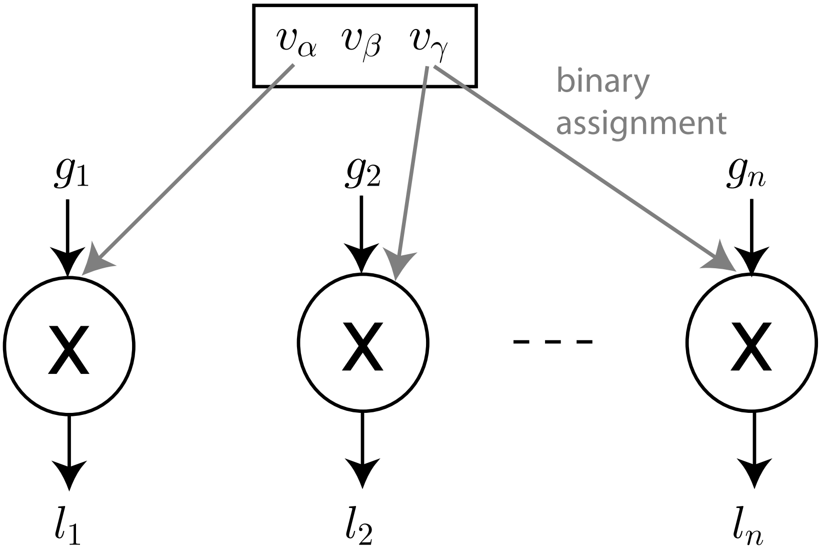


Figure 3: Assignment problem in a multidimensional GSM. Filter responses $\mathbf{l} = \{l_1, \ldots, l_n\}$ are generated by multiplying gaussian variables $\mathbf{g} = \{g_1, \ldots, g_n\}$ by mixer variables $\{v_\alpha, \ldots, v_\mu\}$, where we assume $\mu < n$. We can think of each mixture l_i as the response of a linear filter when applied to a particular image patch. The assignment problem asks which mixer v_j was assigned to each gaussian variable g_i, to form the respective filter response l_i. The set of possible mixers $v_\alpha, v_\beta, v_\gamma$ is surrounded by a rectangular black box. Gray arrows mark the binary assignments: l_1 was generated with mixer v_α, and l_2 and l_n were generated with a common mixer v_γ. In section 4 and Figure 6, we also consider what determines this binary assignment.

demonstration, we assume the standard GSM generative model, in which each gaussian variable is associated with a single mixer variable. In section 4, we extend the GSM generative model to allow probabilistic mixer overlap and propose a solution to the assignment problem. We show that applied to synthetic data, the technique finds the proper assignments and infers correctly the components of the GSM generative model. In section 5, we apply the technique to images. We show that the statistics of the inferred GSM components match the assumptions of the generative model and demonstrate the hierarchical structure that emerges.

3 GSM Inference of Gaussian and Mixer Variables

Consider the simple, single-mixer GSM model described in equation 2.1. We assume $\mathbf{g}$ are uncorrelated, with diagonal covariance matrix $\sigma^2\mathcal{I}$, and that v has a Rayleigh distribution:

$$p[v] \propto \left[v \exp(-v^2/2)\right]^a$$
$$\text{where } 0 < a \leq 1 \text{ parameterizes the strength of the prior.} \tag{3.1}$$

For ease, we develop the theory for $a = 1$. In this case, the variance of each filter response l_i (we will describe the l_i as being filter responses throughout this section, even though they mostly are generated purely synthetically) is exponentially distributed with mean 2. The qualitative properties of the model turn out not to depend strongly on the precise form of $p[v]$. Wainwright et al. (2001) assumed a similar family of mixer variables arising from the square root of a gamma distribution (Wainwright & Simoncelli, 2000), and Portilla et al. considered other forms such as the log normal distribution (Portilla et al., 2001) and a Jeffrey's prior (Portilla et al., 2003).

As stated above, the marginal distribution of the resulting GSM is highly kurtotic (see Figure 1D, left). For our example, given $p[v]$, in fact l follows a double exponential distribution:

$$p[l] \sim \frac{1}{2}\exp(-|l|). \tag{3.2}$$

The joint conditional distribution of two filter responses l_1 and l_2 follows a bowtie shape, with the width of distribution of one response increasing for larger values (both positive and negative) of the other response (see Figure 1D, right).

The inverse problem is to estimate the $n + 1$ variables $g_1, \ldots, g_n, v$ from the n filter responses $l_1, \ldots, l_n$. It is formally ill posed, though regularized through the prior distributions. Four posterior distributions are particularly relevant and can be derived analytically from the model:

1. $p[v|l_1]$ is the local estimate of the mixer, given just a single filter response. In our model, it can be shown that

$$p[v|l_1] = \frac{\sqrt{\frac{\sigma}{|l_1|}}}{\mathcal{B}\left(\frac{1}{2}, \frac{|l_1|}{\sigma}\right)} \exp\left(-\frac{v^2}{2} - \frac{l_1^2}{2v^2\sigma^2}\right), \tag{3.3}$$

where $\mathcal{B}(n, x)$ is the modified Bessel function of the second kind (see also Grenander & Srivastava, 2002). For this, the mean is

$$E[v|l_1] = \sqrt{\frac{|l_1|}{\sigma}} \frac{\mathcal{B}\left(1, \frac{|l_1|}{\sigma}\right)}{\mathcal{B}\left(\frac{1}{2}, \frac{|l_1|}{\sigma}\right)}. \tag{3.4}$$

2. $p[v|\mathbf{l}]$ is the global estimate of the mixer, given all the filter responses. This has a very similar form to $p[v|l_1]$, only substituting $l = \sqrt{\sum_i l_i^2}$ for $|l_1|$,

$$p[v|\mathbf{l}] = \frac{\left(\frac{l}{\sigma}\right)^{\frac{1}{2}(n-2)}}{\mathcal{B}\left(1 - \frac{n}{2}, \frac{l}{\sigma}\right)} v^{-(n-1)} \exp\left(-\frac{v^2}{2} - \frac{l^2}{2v^2\sigma^2}\right), \tag{3.5}$$

whose mean is

$$E[v|\mathbf{l}] = \sqrt{\frac{l}{\sigma}} \frac{\mathcal{B}\left(\frac{3}{2} - \frac{n}{2}, \frac{l}{\sigma}\right)}{\mathcal{B}\left(1 - \frac{n}{2}, \frac{l}{\sigma}\right)}. \tag{3.6}$$

Note that $E[v|\mathbf{l}]$ has also been estimated numerically in noise removal for other mixer variable priors (e.g., Portilla et al., 2001).

3. $p[g_1|l_1]$ is the local estimate of the gaussian variable, given just a local filter response. This is

$$p[g_1|l_1] = \frac{\sqrt{\sigma|l_1|}}{\mathcal{B}\left(-\frac{1}{2}, \frac{|l_1|}{\sigma}\right)} \frac{1}{g_1^2} \exp\left(-\frac{g_1^2}{2\sigma^2} - \frac{l_1^2}{2g_1^2}\right) \mathcal{U}(\text{sign}\{l_1\}g_1), \tag{3.7}$$

where $\mathcal{U}(\text{sign}\{l_1\}g_1)$ is a step function that is 0 if $\text{sign}\{l_1\} \neq \text{sign}\{g_1\}$. The step function arises since g_1 is forced to have the same sign as l_1, as the mixer variables are always positive. The mean is

$$E[g_1|l_1] = \text{sign}\{l_1\}\sigma\sqrt{\frac{|l_1|}{\sigma}} \frac{\mathcal{B}\left(0, \frac{|l_1|}{\sigma}\right)}{\mathcal{B}\left(-\frac{1}{2}, \frac{|l_1|}{\sigma}\right)}. \tag{3.8}$$

4. $p[g_1|\mathbf{l}]$ is the estimate of the gaussian variable, given all the filter responses. Since in our model, the gaussian variables $\mathbf{g}$ are mutually independent, the values of the other filter responses $l_2, \ldots, l_n$ provide information only about the underlying hidden variable v. This leaves $p[g_1|\mathbf{l}]$ proportional to $p(l_1|g_1)P(l_2, \ldots, l_n|v = l_1/g_1)p(g_1)$, which results in

$$p[g_1|\mathbf{l}] = \frac{\sqrt{\sigma|l_1|}\left(\frac{|l_1|}{l}\right)^{\frac{1}{2}(2-n)}}{\mathcal{B}\left(\frac{n}{2} - 1, \frac{l}{\sigma}\right)} g_1^{(n-3)} \exp\left(-\frac{g_1^2}{2\sigma^2}\frac{l^2}{l_1^2} - \frac{l_1^2}{2g_1^2}\right) \mathcal{U}(\text{sign}\{l_1\}g_1) \tag{3.9}$$

with mean

$$E[g_1|\mathbf{l}] = \text{sign}\{l_1\}\sigma\sqrt{\frac{|l_1|}{\sigma}}\sqrt{\frac{|l_1|}{l}} \frac{\mathcal{B}\left(\frac{n}{2} - \frac{1}{2}, \frac{l}{\sigma}\right)}{\mathcal{B}\left(\frac{n}{2} - 1, \frac{l}{\sigma}\right)}. \tag{3.10}$$

We first study inference in this model using synthetic data in which two groups of filter responses $l_1, \ldots, l_{20}$ and $l_{21}, \ldots, l_{40}$ are generated by two mixer variables v_α and v_β (see the schematic in Figure 4A, and the respective

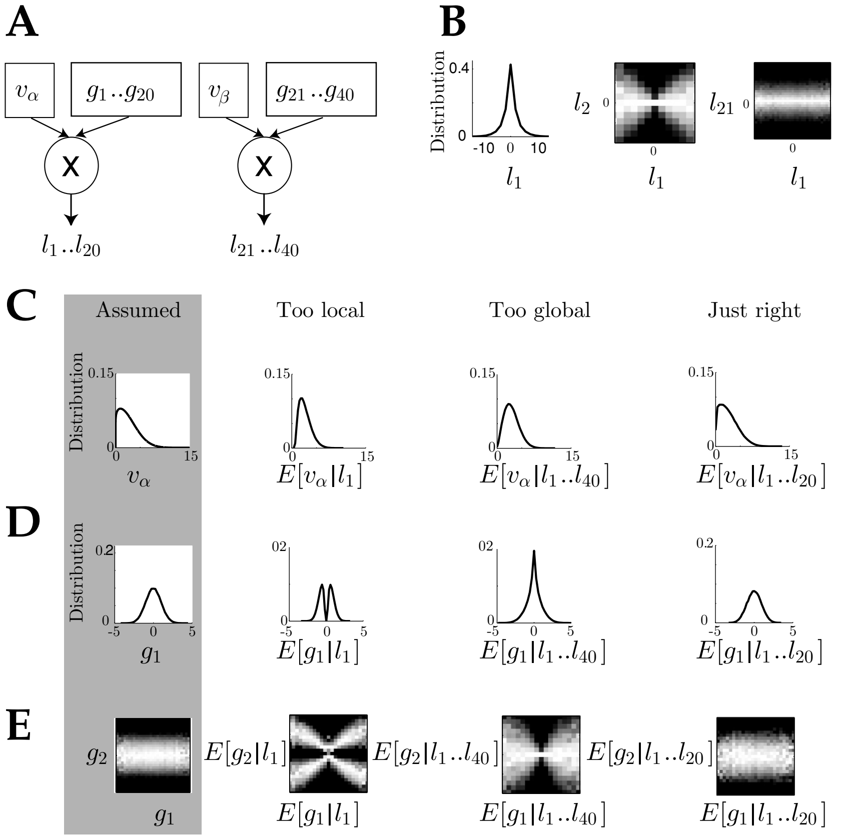


Figure 4: Local and global estimation in synthetic GSM data. (A) Generative model. Each filter response is generated by multiplying its gaussian variable by one of the two mixer variables v_α and v_β. (B) Marginal and joint conditional statistics of sample filter responses. For the joint conditional statistics, intensity is proportional to the bin counts, except that each column is independently rescaled to fill the range of intensities. (C–E) Left column: actual (assumed) distributions of mixer and gaussian variables; other columns: estimates based on different numbers of filter responses (either 1 filter, labeled "too local"; 40 filters, labeled "too global"; or 20 filters, labeled "just right," respectively). (C) Distribution of estimate of the mixer variable v_α. Note that mixer variable values are by definition positive. (D) Distribution of estimate of one of the gaussian variables, g_1. (E) Joint conditional statistics of the estimates of gaussian variables g_1 and g_2.

statistics in Figure 4B). That is, each filter response is deterministically generated from either mixer v_α or mixer v_β, but not both. We attempt to infer the gaussian and mixer components of the GSM model from the synthetic data, assuming that we do not know the actual mixer assignments.

Figures 4C and 4D show the empirical distributions of estimates of the conditional means of a mixer variable $E(v_\alpha|\{l\})$ (see equations 3.4 and 3.6) and one of the gaussian variables $E(g_1|\{l\})$ (see equations 3.8 and 3.10) based on different assumed assignments. For inference based on too few filter responses, the estimates do not match the actual distributions (see the second column labeled "too local"). For example, for a local estimate based on a single filter response, the gaussian estimate peaks away from zero. This is because the filter response is a product of the two terms, the gaussian and the mixer, and the problem is ill posed with only a single filter estimate. Similarly, the mixer variable is not estimated correctly for this local case. Note that this occurs even though we assume the correct priors for both the mixer and gaussian variables and is thus a consequence of the incorrect assumption about the assignments. Inference is also compromised if it is based on too many filter responses, including those generated by both v_α and v_β (see the third column, labeled "too global"). This is because inference of v_α is based partly on data that were generated with a different mixer, v_β (so when one mixer is high, the other might be low, and so on). In contrast, if the assignments are correct and inference is based on all those filter responses that share the same common generative mixer (in this case v_α), the estimates become good (see the last column, labeled "just right").

In Figure 4E, we show the joint conditional statistics of two components, each estimating their respective g_1 and g_2. Again, as the number of filter responses increases, the estimates improve, provided that they are taken from the right group of filter responses with the same mixer variable v_α. Specifically, the mean estimates of g_1 and g_2 become more independent (see the last column). Note that for estimations based on a single filter response, the joint conditional distribution of the gaussian appears correlated rather than independent (second column); for estimation based on too many filter responses generated from either of the mixer variables, the joint conditional distribution of the gaussian estimates shows a dependent (rather than independent) bowtie shape (see the third column). Mixer variable joint statistics also deviate from their actual independent forms when the estimations are too local or global (not shown). These examples indicate modes of estimation failure for synthetic GSM data if one does not know the proper assignments between mixer and gaussian variables. This suggests the need to infer the appropriate assignments from the data.

To show that this is not just a consequence of an artificial example, we consider estimation for natural image data. Figure 5 demonstrates estimation of mixer and gaussian variables for an example natural image. We derived linear filters from a multiscale oriented steerable pyramid (Simoncelli, Freeman, Adelson, & Heeger, 1992), with 100 filters, at two preferred

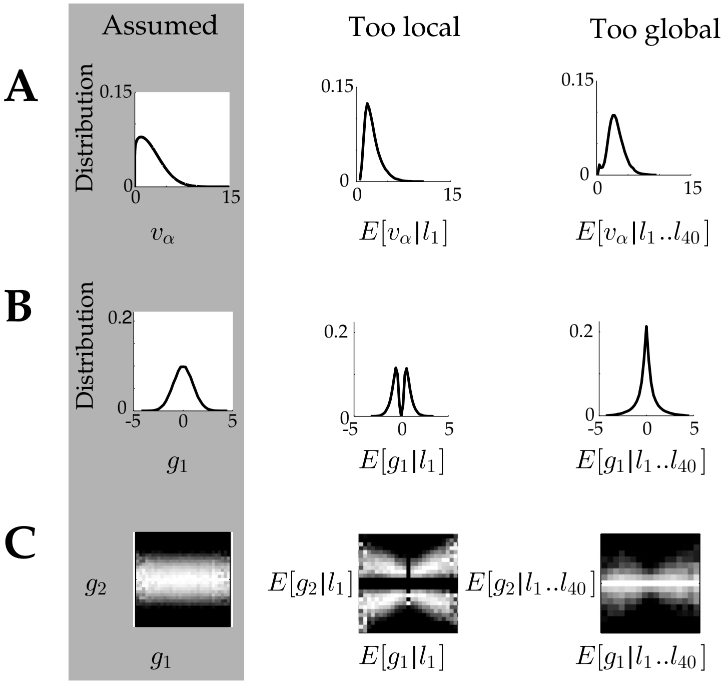


Figure 5: Local and global estimation in image data. (A–C) Left: Assumed distributions of mixer and gaussian variables; other columns: estimates based on different numbers of filter responses (either 1 filter, labeled "too local," or 40 filters, including two orientations across a 38 by 38 pixel region, labeled "too global," respectively). (A) Distribution of estimate of the mixer variable v_α. Note that mixer variable values are by definition positive. (B) Distribution of estimate of one of the gaussian variables, g_1. (C) Joint conditional statistics of the estimates of gaussian variables g_1 and g_2.

orientations, 25 nonoverlapping spatial positions (with spatial subsampling of 8 pixels), and a single phase and spatial frequency peaked at 1/6 cycles per pixel. By fitting the marginal statistics of single filters, we set the Rayleigh parameter of equation 3.1 to $a = 0.1$. Since we do not know a priori the actual assignments that generated the image data, we demonstrate examples for which inference is either very local (based on a single wavelet coefficient input) or very global (based on 40 wavelet coefficients at two orientations and a range of spatial positions).

Figure 5 shows the inferred marginal and bowtie statistics for the various cases. If we compare the second and third columns to the equivalents in Figures 4C to 4E for the synthetic case, we can see close similarities. For instance, overly local or global inference of the gaussian variable leads to bimodal or leptokurtotic marginals, respectively. The bowtie plots are also similar. Indeed, we and others (Ruderman & Bialek, 1994; Portilla et al., 2003) have observed changes in image statistics as a function of the width of the spatial neighborhood or the set of wavelet coefficients.

It would be ideal to have a column in Figure 5 equivalent to the "just right" column of Figure 4. The trouble is that the equivalent neighborhood of a filter is defined not merely by its spatial extent, but rather by all of its featural characteristics and in an image and image-class dependent manner. For example, we might expect different filter neighborhoods for patches with a vertical texture everywhere than for patches corresponding to an edge or to features of a face. Thus, different degrees of local and global arrangements may be appropriate for different images. Since we do not know how to specify the mixer groups a priori, it is desirable to learn the assignments from a set of image samples. Furthermore, it may be necessary to have a battery of possible mixer groupings available to accommodate the statistics of different images.

4 Solving the Assignment Problem

The plots in Figures 4 and 5 suggest that it should be possible to infer the assignments, that is, work out which linear filters share common mixers, by learning from the statistics of the resulting joint dependencies. Further, real-world stimuli are likely better captured by the possibility that inputs are coordinated in somewhat different collections in different images. Hard assignment problems, in which each input pays allegiance to just one mixer, are notoriously computationally brittle. Soft assignment problems, in which there is a probabilistic relationship between inputs and mixers, are computationally better behaved. We describe the soft assignment problem and illustrate examples with synthetic data. In section 5, we turn to image data.

Consider the richer mixture-generative GSM shown in Figure 6. To model the generation of filter responses l_i for a single image patch (see Figure 6A), we multiply each gaussian variable g_i by a single mixer variable from the set $v_\alpha, v_\beta, \ldots, v_\mu$. In the deterministic (hard assignment) case, each gaussian variable is associated with a fixed mixer variable in the set. In the probabilistic (soft assignment) case, we assume that g_i has association probability p_{ij} (satisfying $\sum_j p_{ij} = 1, \forall i$) of being assigned to mixer variable v_j. Note that this is a competitive process, by which only a single mixer variable is assigned to each filter response l_i in each patch, and the assignment is determined according to the association probabilities. As a result, different image patches will have different assignments (see Figures 6A and 6B). For example, an image patch with strong vertical texture everywhere might have quite different assignments from an image patch with a vertical edge on the right corner. Consequently, in these two patches, the linear filters will share different common mixers. The assignments are assumed to be made independently for each patch. Therefore, the task for hierarchical learning is to work out association probabilities suitable for generating the filter responses. We use $\chi_i \in \{\alpha, \beta, \ldots \mu\}$ for the assignments

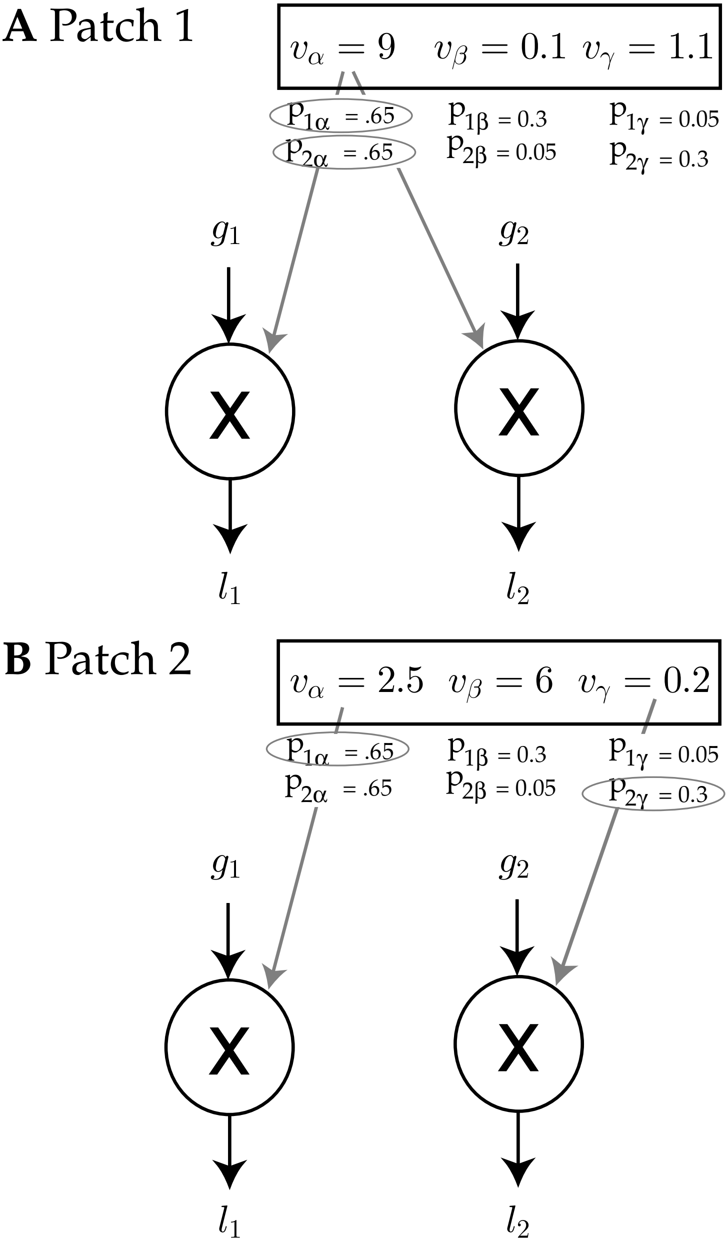
$$l_i = g_i v_{\chi_i}. \tag{4.1}$$

Consider a specific synthetic example of a soft assignment: 100 filter responses are generated probabilistically from three mixer variables, v_α, v_β, and v_γ. Figure 7A shows the association probabilities p_{ij}. Figure 8A shows example marginal and joint conditional statistics for the filter responses, based on an empirical sample of 5000 points drawn from the generative model. On the left is the typical bowtie shape between two filter responses generated with the same mixer, v_α, 100% of the time. In the middle is a weaker dependency between two filter responses whose mixers overlap for only some samples. On the right is an independent joint conditional distribution arising from two filter responses whose mixer assignments do not overlap.

There are various ways to try solving soft assignment problems (see, e.g., MacKay, 2003). Here we use the Markov chain Monte Carlo method called Gibbs sampling. The advantage of this method is its flexibility and power. Its disadvantage is its computational expense and biological implausibility—although for the latter, we should stress that we are mostly interested in an abstract characterization of the higher-order dependencies rather than in a model for activity-dependent representation formation. Williams and Adams (1999) suggested using Gibbs sampling to solve a similar assignment problem in the context of dynamic tree models. Variational approximations have also been considered in this context (Adams & Williams, 2003; Hinton et al., 1999).

Inference and learning in this model proceeds in two stages, according to an expectation maximization framework (Dempster et al., 1977). First, given a filter response l_i, we use Gibbs sampling to find possible

Figure 6: Extended generative GSM model with soft assignment. (A) The depiction is similar to Figure 3, except that we examine only the generation of two of the filter responses l_1 and l_2, and we show the probabilistic process according to which the assignments are made. The mixer variable assigned to l_1 is chosen for each image patch according to the association probabilities $p_{1\alpha}$, $p_{1\beta}$, and $p_{1\gamma}$. The binary assignment for filter response l_1 corresponds to mixer $v_\alpha = 9$. The binary choice arose from the higher association probability $p_{1\alpha} = 0.65$, marked with a gray ellipse. The assignment is marked by a gray arrow. For this patch, the assignment for filter l_2 also corresponds to $v_\alpha = 9$. Thus, l_1 and l_2 share a common mixer (with a relatively high value). (B) The same for a second patch; here assignment for l_1 corresponds to $v_\alpha = 2.5$, but for l_2 to $v_\gamma = 0.2$.


A Patch 1
$v_\alpha = 9$ $v_\beta = 0.1$ $v_\gamma = 1.1$
$p_{1\alpha} = .65$
$p_{2\alpha} = .65$
$p_{1\beta} = 0.3$
$p_{2\beta} = 0.05$
$p_{1\gamma} = 0.05$
$p_{2\gamma} = 0.3$
g_1
g_2
X
X
l_1
l_2
B Patch 2
$v_\alpha = 2.5$ $v_\beta = 6$ $v_\gamma = 0.2$
$p_{1\alpha} = .65$
$p_{2\alpha} = .65$
$p_{1\beta} = 0.3$
$p_{2\beta} = 0.05$
$p_{1\gamma} = 0.05$
$p_{2\gamma} = 0.3$
g_1
g_2
X
X
l_1
l_2

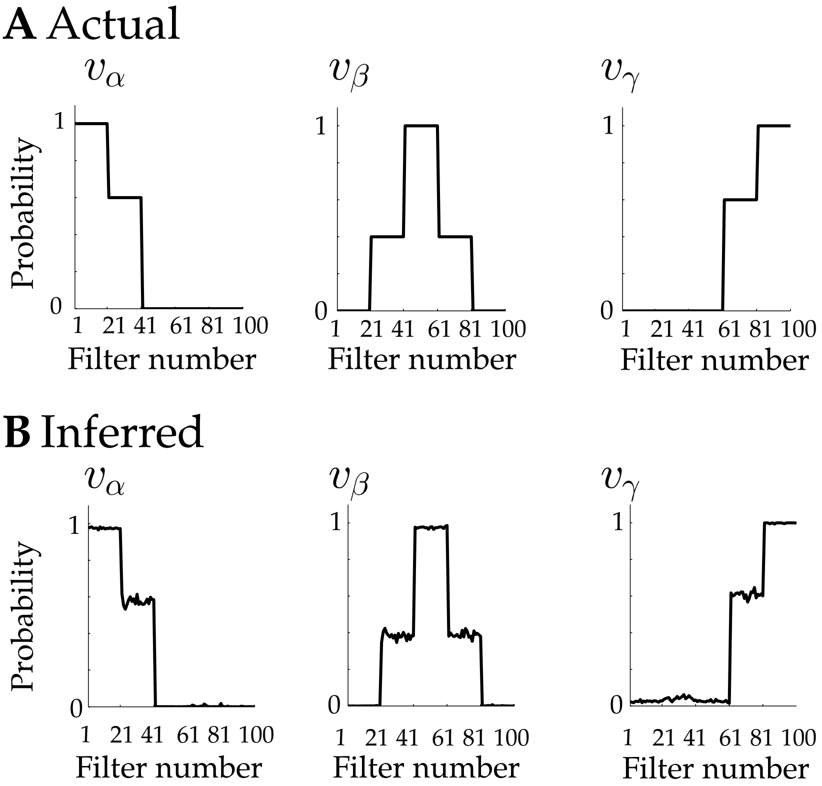


Figure 7: Inference of mixer association probabilities in a synthetic example. (A) Each filter response l_i is generated by multiplying its gaussian variable by a probabilistically chosen mixer variable, v_α, v_β, or v_γ. Shown are the actual association probabilities p_{ij} (labeled probability) of the generated filter responses l_i with each of the mixer variables v_j. (B) Inferred association probabilities p_{ij} from the Gibbs procedure, corresponding to v_α, v_β, and v_γ.

appropriate (posterior) assignments to the mixers. Second, given the collection of assignments across multiple filter responses, we update the association probabilities p_{ij} (see the appendix).

We tested the ability of this inference method to find the association probabilities in the overlapping mixer variable synthetic example shown in Figure 7A. The Gibbs sampling procedure requires that we specify the number of mixer variables that generated the data. In the synthetic example, the actual number of mixer variables is 3. We ran the Gibbs sampling procedure, assuming the number of possible mixer variables is 5 (e.g., > 3). After 500 iterations, the weights converged near the proper probabilities. In

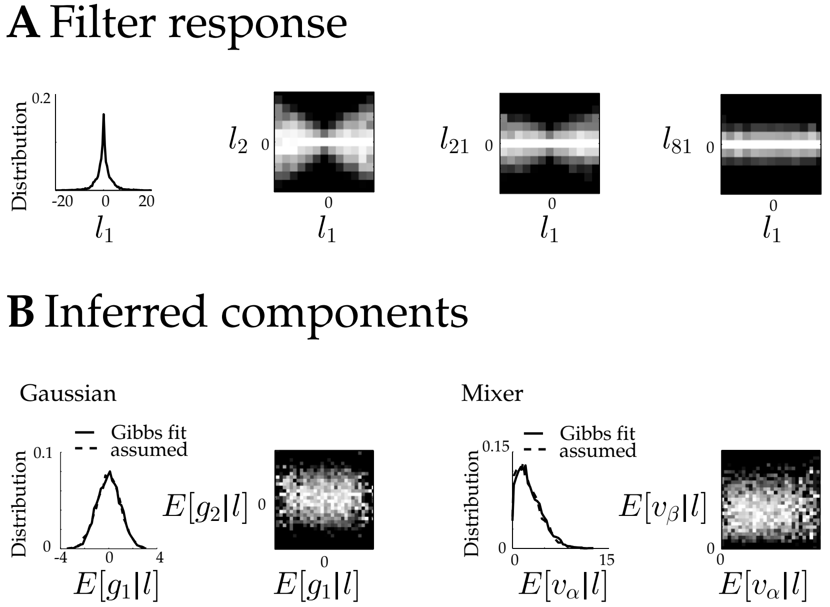


Figure 8: Inference of gaussian and mixer components in a synthetic example. (A) Example marginal and joint conditional filter response statistics. (B) Statistics of gaussian and mixer estimates from Gibbs sampling.

Figure 7A, we plot the actual probability distributions for the filter response associations with each of the mixer variables. In Figure 7B, we show the estimated associations for three of the mixers: the estimates closely match the actual association probabilities; the other two estimates yield association probabilities near zero, as expected (not shown).

We estimated the gaussian and mixer components of the GSM using the Bayesian equations of the previous section (see equations 3.10 and 3.6), but restricting the input samples to those assigned to each mixer variable. In Figure 8B, we show examples of the estimated distributions of the gaussian and mixer components of the GSM. Note that the joint conditional statistics of both gaussian and mixer are independent, since the variables were generated as such in the synthetic example. The Gibbs procedure can be adjusted for data generated with different Rayleigh parameters *a* (in equation 3.1), allowing us to model a wide range of behaviors observed in the responses of linear filters to a range of images. We have also tested the synthetic model for cases in which the mixer variable generating the data deviates somewhat from the assumed mixer variable distribution: Gibbs sampling still tends to find the proper association weights, but the probability distribution estimate of the mixer random variable is not matched to the assumed distribution.

We have thus far discussed the association probabilities determined by Gibbs inference for filter responses over the full set of patches. How does Gibbs inference choose the assignments on a patch-by-patch basis? For filter responses generated deterministically, according to a single mixer, the learned association probabilities of filter responses to this mixer are approximately equal to a probability of 1, and so the Gibbs assignments are correct approximately 100% of the time. For filter responses generated probabilistically from more than one mixer variable (e.g., filter responses 21–40 or 61–80 for the example in Figures 7 and 8), there is potential ambiguity about the generating mixer. We focus specifically on filter responses 21 to 40, which are generated from either v_α or v_β. Note that the overall association probabilities for the mixers for all patches are 0.6 and 0.4, respectively. We would like to know how these are distributed on a patch-by-patch basis.

To assess the correctness of the Gibbs assignments, we repeated 40 independent runs of Gibbs sampling for the same filter responses and computed the percentage correct assignment for filter responses that were generated according to v_α or v_β (note that we know the actual generating mixer values for the synthetic data). We did this on a patch-by-patch basis and found that two factors affected the Gibbs inference: (1) the ratio of the two mixer variables v_β/v_α for the given patch and (2) the absolute value of the ambiguous filter response for the given patch. Figure 9 summarizes the Gibbs assignments. The x-axis indicates the ratio of the absolute value of the ambiguous filter response and v_α. The y-axis indicates the percentage correct for filter responses that were actually generated from v_α (black circles) or v_β (gray triangles). In Figure 9A we depict the result for a patch in which the ratio v_β/v_α was approximately 1/10 (marked by an asterisk on the x-axis). This indicates that filter responses generated by v_α are usually larger than filter responses generated by v_β, and so for sufficiently large or small (absolute) filter response values, it should be possible to determine the generating mixer. Indeed, Gibbs assigns correctly filter responses for which the ratio of the filter response and v_α are reasonably above or below 1/10 but does not fare as well for ratios that are in the range of 1/10 and could have potentially been generated by either mixer. Figure 9B illustrates a similar result for $v_\beta/v_\alpha \approx 1/3$. Finally, Figure 9C shows that for $v_\beta/v_\alpha \approx 1$, all filter responses are in the same range, and Gibbs resorts to the approximate association probabilities, of 0.6 and 0.4, respectively.

We also tested Gibbs inference in undercomplete cases for which the Gibbs procedure assumes fewer mixer variables than were actually used to generate the data. Figure 10 shows an example in which we generated 75 sets of filter responses according to 15 mixer variables, each associated deterministically with five (consecutive) filter responses. We ran Gibbs assuming that only 10 mixers were collectively responsible for all the filter responses. Figure 10 shows the actual and inferred association probabilities

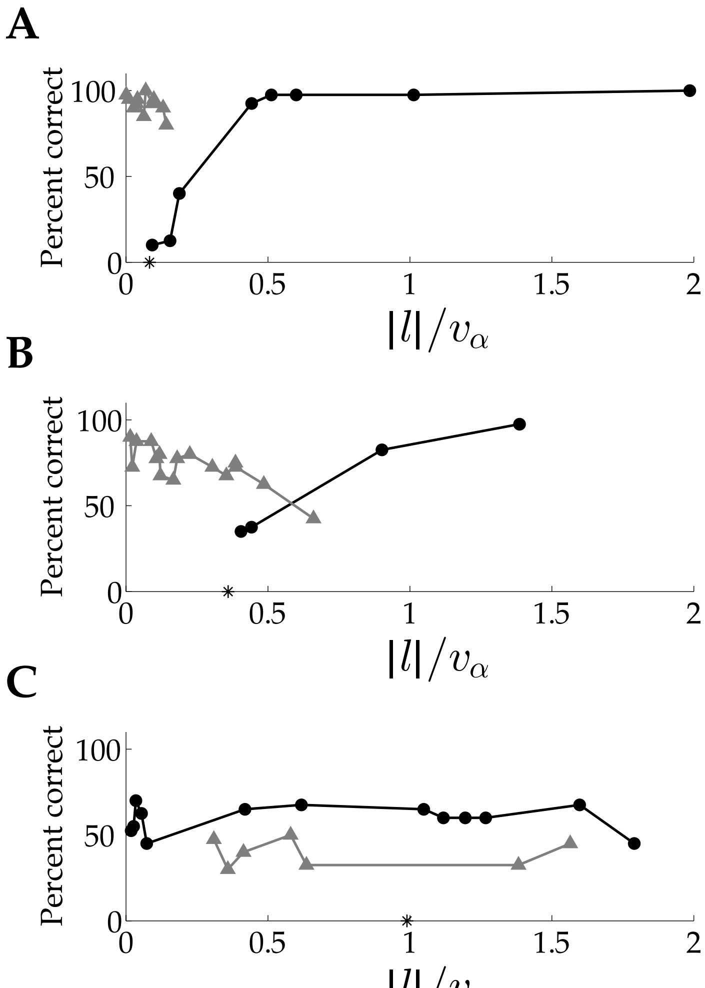


Figure 9: Gibbs assignments on a patch-by-patch basis in a synthetic example. For filter responses of each patch (here, filter responses 21–40), there is ambiguity as to whether the assignments were generated according to v_α or v_β (see association probabilities in Figure 7). We summarize the percentage correct assignments computed over 40 independent Gibbs runs (y-axis), separately for the patches with filter responses actually generated according to v_α (black, circles) and filter responses actually generated according to v_β (gray, triangles). There are overall 20 points corresponding to the 20 filter responses. For readability, we have reordered v_α and v_β, such that $v_\alpha \geq v_\beta$. The x-axis depicts the ratio of the absolute value of each ambiguous filter response in the patch (labeled "patch"), and v_α. The black asterisk on the x-axis indicates the ratio v_β/v_α. See the text for interpretation. (A) $v_\beta/v_\alpha \approx 1/10$. (B) $v_\beta/v_\alpha \approx 1/3$. (C) $v_\beta/v_\alpha \approx 1$.

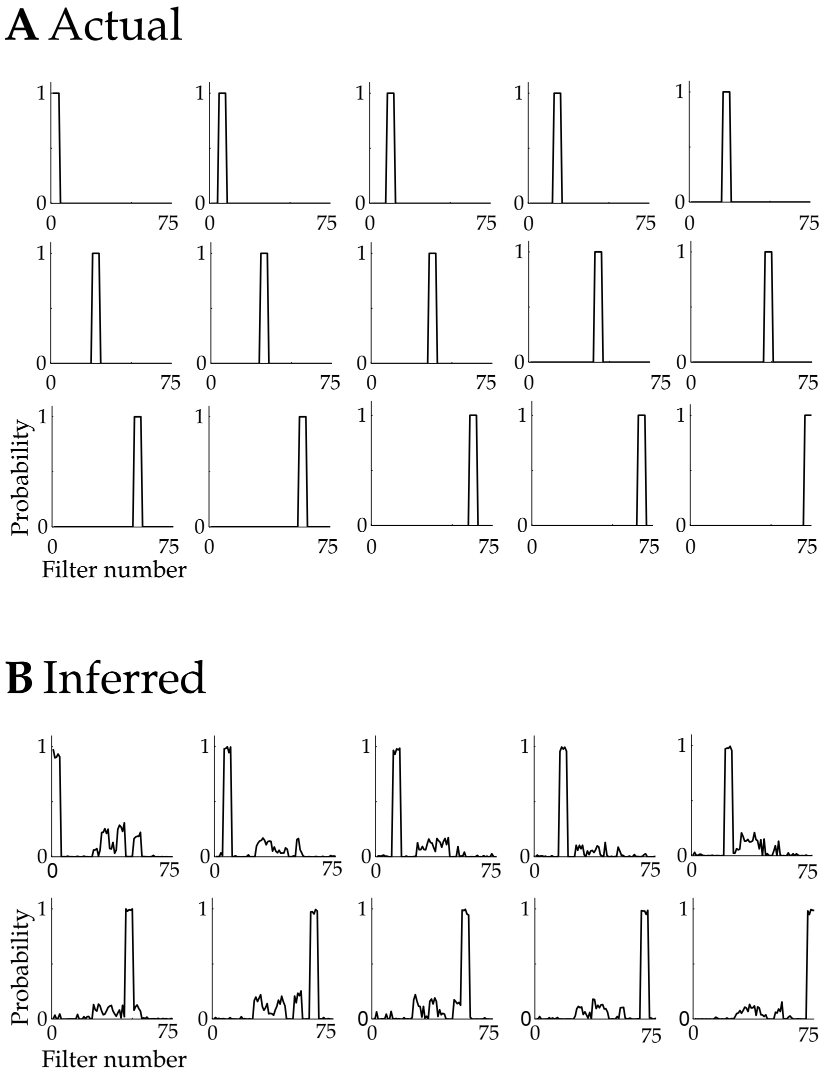


Figure 10: Inference of mixer association probabilities in an undercomplete synthetic example. The data were synthesized with 15 mixer variables, but Gibbs inference assumes only 10 mixer variables. (A) Actual association probabilities. Note that assignments are deterministic, with 0 or 1 probability, in consecutive groups of 5. (B) Inferred association probabilities.

in this case. The procedure correctly groups together five filters in each of the 10 inferred associations. There are groups of five filters that are not represented by a single high-order association, and these are spread across the other associations, with smaller weights. The added noise is expected, since the association probabilities for each filter must sum to 1.

5 Image Data

Having validated the inference model using synthetic data, we turned to natural images. Here, the l_i are actual filter responses rather than synthesized products of a generative model. We considered inference on both wavelet filters and ICA bases and with a number of different image sets.

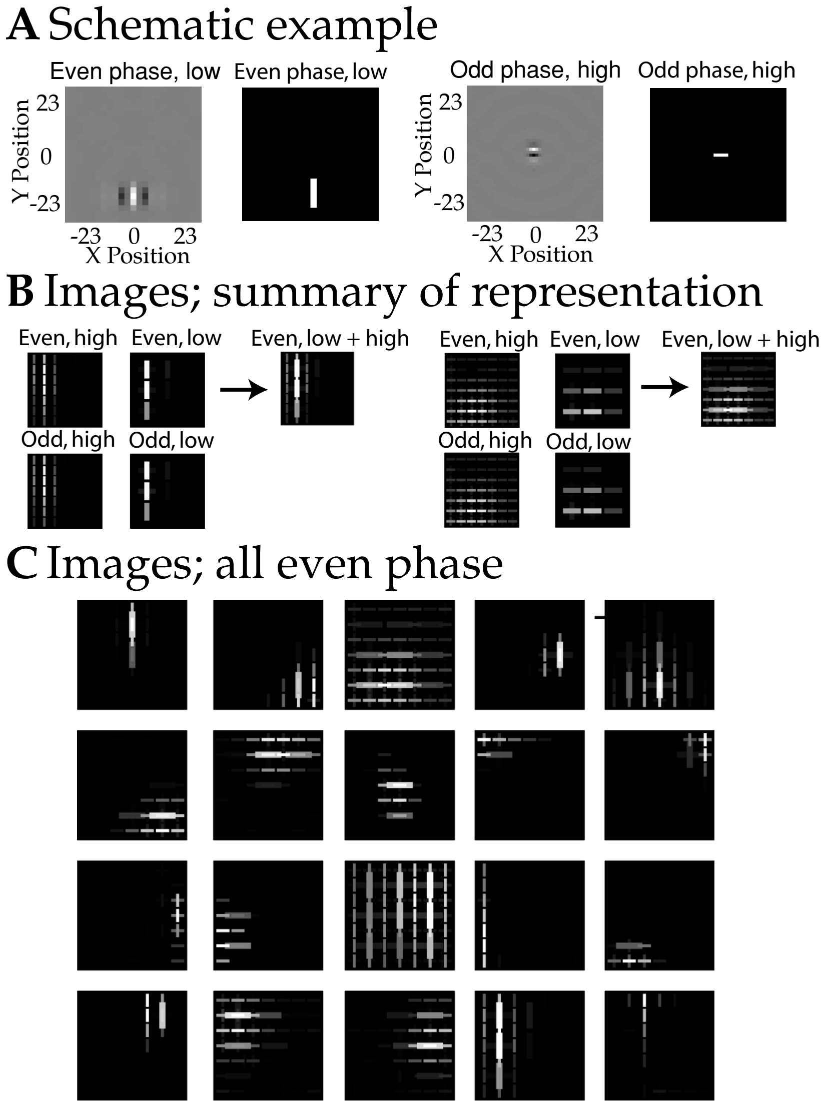
We first derived linear filters from a multiscale oriented steerable pyramid (Simoncelli et al., 1992), with 232 filters. These consist of two phases (even and odd quadrature pairs), two orientations, and two spatial frequencies. The high spatial frequency is peaked at approximately 1/6 cycles per pixel and consists of 49 nonoverlapping spatial positions. The low spatial frequency is peaked at approximately 1/12 cycles per pixel, and consists of 9 nonoverlapping spatial positions. The spacing between filters, along vertical and horizontal spatial shifts, is 7 pixels (higher frequency) and 14 pixels (lower frequency). We used an ensemble of five images from a standard compression database (see Figure 12A) and 8000 samples.

We ran our method with the same parameters as for synthetic data, with 20 possible mixer variables and Rayleigh parameter $a = 0.1$. Figure 11 shows the association probabilities p_{ij} of the filter responses for each of the obtained mixer variables. In Figure 11A, we show a schematic (template) of the association representation that follows in Figure 11B for the actual data. Each set of association probabilities for each mixer variable is shown for coefficients of two phases, two orientations, two spatial frequencies, and the range of spatial positions along the vertical and horizontal axes. Unlike the synthetic examples, where we plotted the association probabilities in one dimension, for the images we plot the association probabilities along a two-dimensional spatial grid matched to the filter set.

We now study the pattern of the association probablilities for the mixer variables. For a given mixer, the association probabilities signify the probability that filter responses were generated with that mixer. If a given mixer variable has high association probabilities corresponding to a particular set of filters, we say that the mixer neighborhood groups together the set of filters. For instance, the mixer association probabilities in Figure 11B (left) depict a mixer neighborhood that groups together mostly vertical filters on the left-hand side of the spatial grid, of both even and odd phase. Strikingly, all of the mixer neighborhoods group together two phases of quadrature pair. Quadrature pairs have also been extracted from cortical data (Touryan, Lau, & Dan, 2002; Rust, Schwartz, Movshon, & Simoncelli, 2005) and are the components of ideal complex cell models. However, the range of spatial groupings of quadrature pairs that we obtain here has not been reported in visual cortex and thus constitutes a prediction of the model. The mixer neighborhoods range from sparse grouping across space to more global grouping. Single orientations are often grouped across space, but in a couple of cases, both orientations are grouped together. In addition, note that there is some probabilistic overlap between mixer neighborhoods; for instance, the global vertical neighborhood associated with one of the mixers overlaps with other more localized, vertical neighborhoods associated with other mixers. The diversity of mixer neighborhoods matches our intuition that different mixer arrangements may be appropriate for different image patches.

We examine the image patches that maximally activate the mixers, similar to Karklin and Lewicki (2003a). In Figure 12 we show different mixer association probabilities and patches with the maximum log likelihood of $P(v|patch)$. One example mixer neighborhood (see Figure 12B) is associated with global vertical structure across most of its "receptive" region. Consequently, the maximal patches correspond to regions in the image data with multiple vertical structure. Another mixer neighborhood (see Figure 12C) is associated with vertical structure in a more localized iso-oriented region of space; this is also reflected in the maximal patches. This is perhaps similar to contour structure that is reported from the statistics of natural scenes (Geisler, Perry, Super, & Gallogly, 2001; Hoyer & Hyvärinen, 2002). Another mixer neighborhood (see Figure 12D) is associated with vertical and horizontal structure in the corner, with maximal patches that tend to have any structure in this region (a roof corner, an eye, a distant face, and so on). The mixer neighborhoods in Figures 12B and 12D bear similarity to those in Karklin and Lewicki (2003a).

Figure 11: Inference of mixer association probabilities for images and wavelet filters. (A) Schematic of filters and association probabilities for a single mixer, on a 46-by-46 pixel spatial grid (separate grid for even and odd phase filters). Left: Example even phase filter along the spatial grid. To the immediate right are the association probabilities. The probability that each filter response is associated with the mixer variable ranges from 0 (black) to 1 (white). Only the example filter has high probability, in white, with a vertical line representing orientation. Right: Example odd phase filter and association probabilities (the small line represents higher spatial frequency). (B) Example mixer association probabilities for image data. Even and odd phases always show a similar pattern of probabilities, so we summarize only the even phase probability and merge together the low- and high-frequency respresentation. (C) All 20 mixer association probabilities for image data for the even phase (arbitrarily ordered). Each probability plot is separately normalized to cover the full range of intensities.


A Schematic example
Even phase, low
Even phase, low
Odd phase, high
Odd phase, high
Y Position
23
0
-23
-23
0
23
X Position
Y Position
23
0
-23
-23
0
23
X Position
B Images; summary of representation
Even, high
Even, low
Even, low + high
Odd, high
Odd, low
Even, high
Even, low
Even, low + high
Odd, high
Odd, low
C Images; all even phase

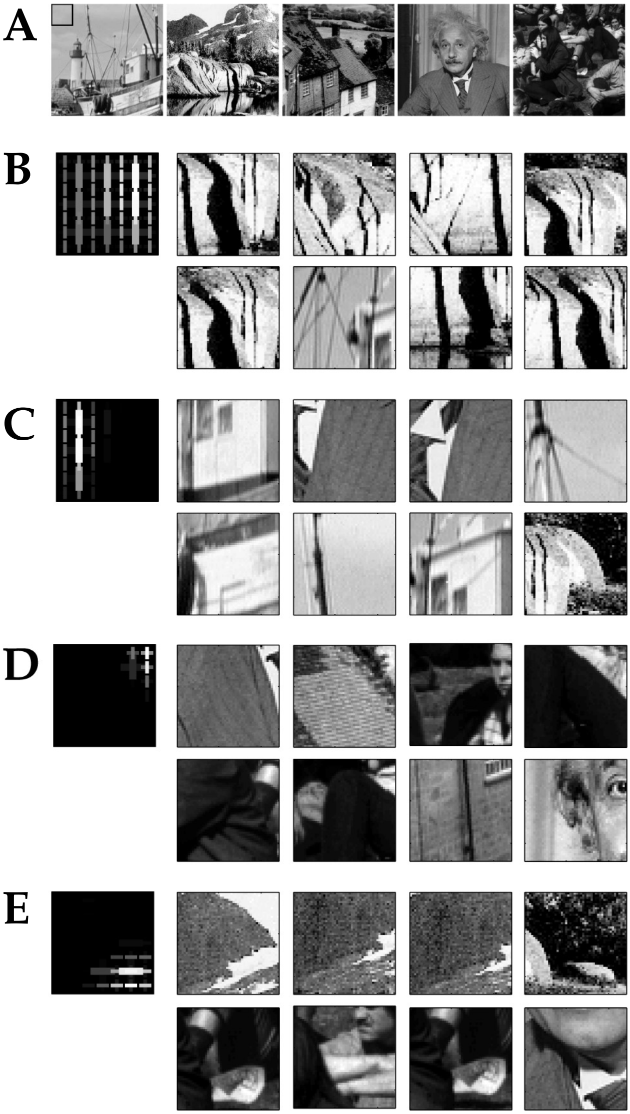


Figure 12: Maximal patches for images and wavelet filters. (A) Image ensemble. Black box marks the size of each image patch. (B–E) Example mixer association probabilities and 46×46 pixel patches that had the maximum log likelihood of $P(v|patch)$.

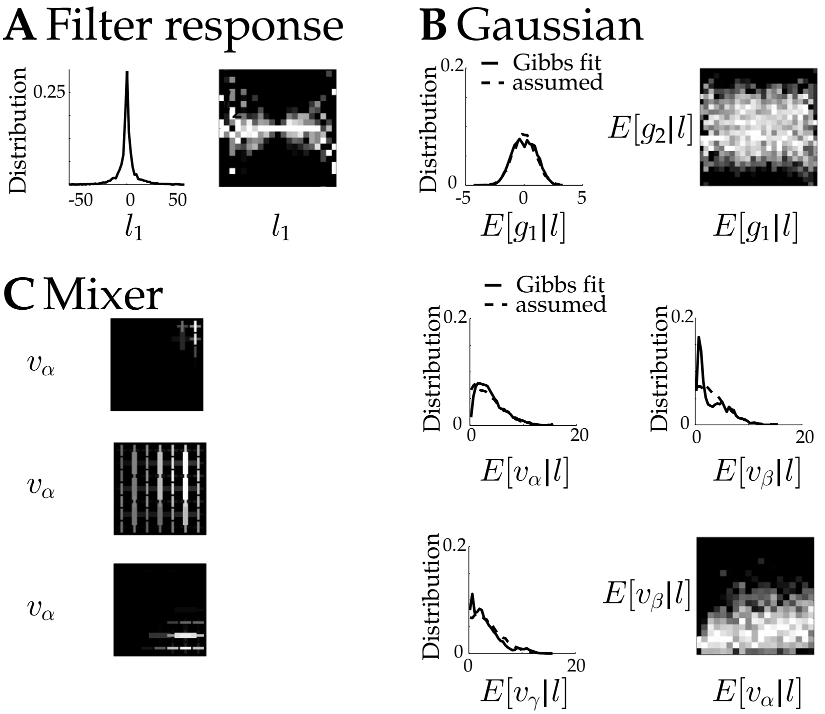


Figure 13: Inference of gaussian and mixer components for images and wavelet filters. (A) Statistics of images through one filter and joint conditional statistics through two filters. Filters are quadrature pairs, spatially displaced by seven pixels. (B) Inferred gaussian statistics following Gibbs. The dashed line is assumed statistics, and the solid line is inferred statistics. (C) Statistics of example inferred mixer variables following Gibbs. On the left are the mixer association probabilities, and the statistics are shown on the right.

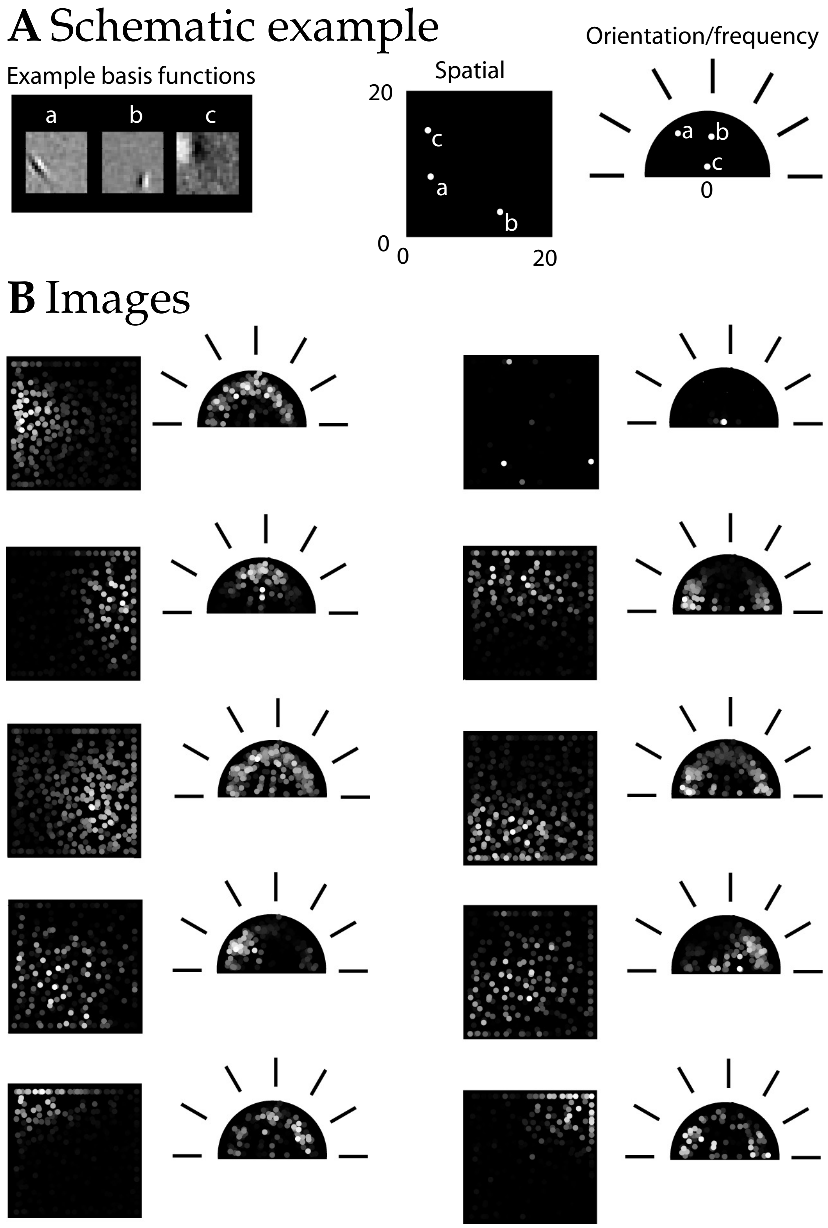
Although some of the mixer neighborhoods have a localized responsive region, it should be noted that they are not sensitive to the exact phase of the image data within their receptive region. For example, in Figure 12C, it is clear that the maximal patches are invariant to phase. This is to be expected, given that the neighborhoods are always arranged in quadrature pairs.

From these learned associations, we also used our model to estimate the gaussian and mixer variables (see equations 3.10 and 3.6). In Figure 13, we show representative statistics for the filter responses and the inferred variables. The learned distributions of gaussian and mixer variables match our assumptions reasonably well. The gaussian estimates exhibit joint conditional statistics that are roughly independent. The mixer variables are typically (weakly) dependent.

To test if the result is not merely a consequence of the choice of wavelet-based linear filters and natural image ensemble, we ran our method on the responses of filters that arose from ICA (Olshausen & Field, 1996) and with 20-by-20 pixel patches from Field's image set (Field, 1994; Olshausen & Field, 1996). Figure 14 shows example mixer neighborhood associations in terms of the spatial and orientation/frequency profile and corresponding weights (Karklin & Lewicki, 2003a). The association grouping consists of both spatially global examples that group together a single orientation at all spatial positions and frequencies and more localized spatial groupings. The localized spatial groupings sometimes consist of all orientations and spatial frequencies (as in Karklin & Lewicki, 2003a) and are sometimes more localized in these properties (e.g., a vertical spatial grouping may tend to have large weights associated with roughly vertical filters). The statistical properties of the components are similar to the wavelet example (not shown here). Example maximal patches are shown in Figure 15. In Figure 15B are maximal patches associated with a spatially global diagonal structure; in Figure 15C are maximal patches associated with approximately vertical orientation on the right-hand side; in Figure 15D are maximal patches associated with low spatial frequencies. Note that there is some similarity to Karklin and Lewicki (2003a) in the maximal patches.

So far we have demonstrated inference for a heterogeneous ensemble of images. However, it is also interesting and perhaps more intuitive to consider inference for particular images or image classes. We consider a couple of examples with wavelet filters, in which we both learn and demonstrate the results on the particular image class. In Figure 16 we demonstrate example mixer association probabilities that are learned for a zebra image (from a Corel CD-ROM). As before, the neighborhoods are composed of quadrature pairs (only even phase shown); however, some of the spatial configurations are richer. For example, in Figure 16A, the mixer neighborhood captures a horizontal-bottom/vertical-top spatial configuration. In Figure 16B, the mixer neighborhood captures a global vertical configuration, largely present in the back zebra, but also in a

Figure 14: Inference of mixer association probabilities for Field's image ensemble (Field, 1994) and ICA bases. (A) Schematic example of the representation for three basis functions. In the spatial plot, each point is the center spatial location of the corresponding basis function. In the orientation/frequency plot, each point is shown in polar coordinates where the angle is the orientation and the radius is the frequency of the corresponding basis function. (B) Example mixer association probabilities learned from the images. Each probability plot is separately normalized to cover the full range of intensities.


A Schematic example
Example basis functions
a
b
c
Spatial
20
0
0
20
c
a
b
Orientation/frequency
a
b
c
0
B Images

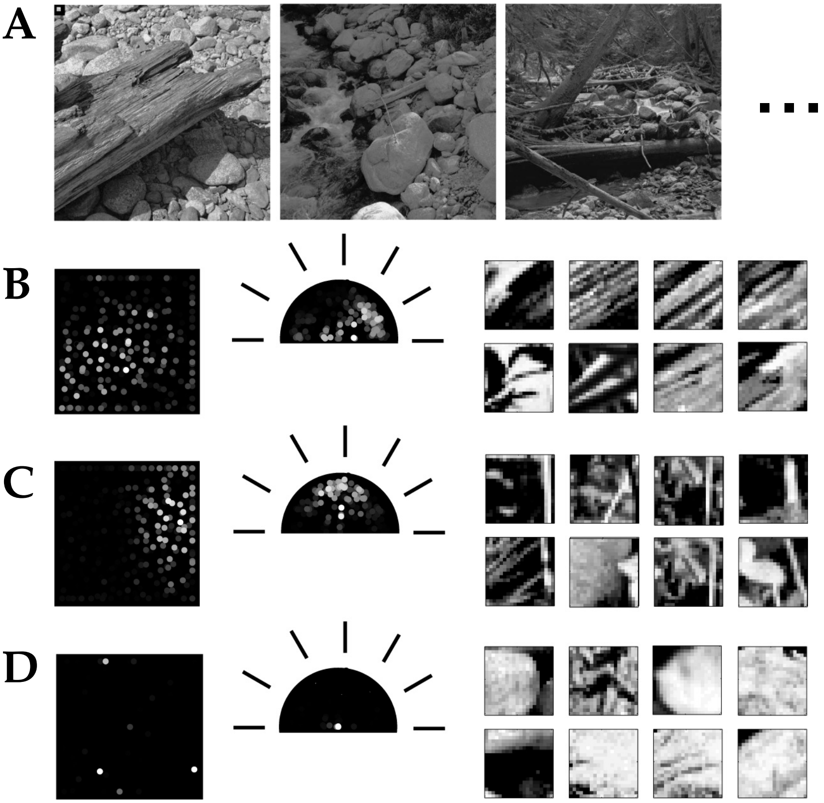


Figure 15: Maximal patches for Field's image ensemble (Field, 1994) and ICA bases. (A) Example input images. The black box marks the size of each image patch. (B–D) The 20×20 pixel patches that had the maximum log likelihood of $P(v|patch)$.

portion of the front zebra. Some neighborhoods (not shown here) are more local.

We also ran Gibbs inference on a set of 40 face images (20 different people, 2 images of each) (Samaria & Harter, 1994). The mixer neighborhoods are again in quadrature pairs (only even phase shown). Some of the more interesting neighborhoods appear to capture richer information that is not necessarily continuous across space. Figure 17A shows a neighborhood resembling a loose sketch of the eyes, the nose, and the mouth (or moustache); the maximal patches are often roughly centered accordingly. The neighborhood in Figure 17B is also quite global but more abstract and appears to largely capture the left edge of the face along with other features. Figure 17C

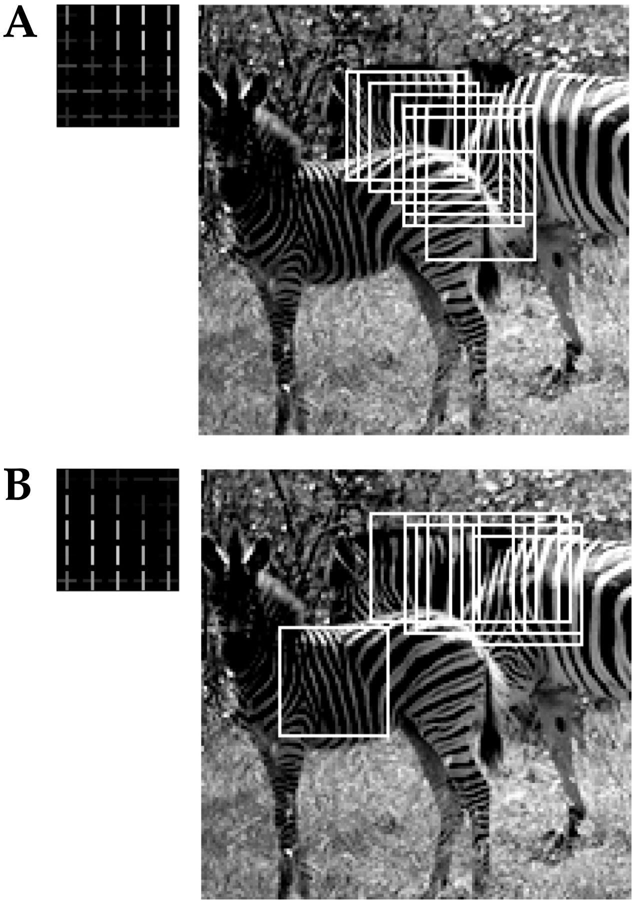


Figure 16: (A–B) Example mixer association probabilities and maximal patches for zebra image and wavelets. Maximal patches are marked with white boxes on the image.

shows a typical local neighborhood, which captures features within its receptive region but is rather nonspecific.

6 Discussion

The study of natural image statistics is evolving from a focus on issues about scale-space hierarchies and wavelet-like components and toward the coordinated statistical structure of the wavelets. Bottom-up ideas (e.g., bowties, hierarchical representations such as complex cells) and top-down

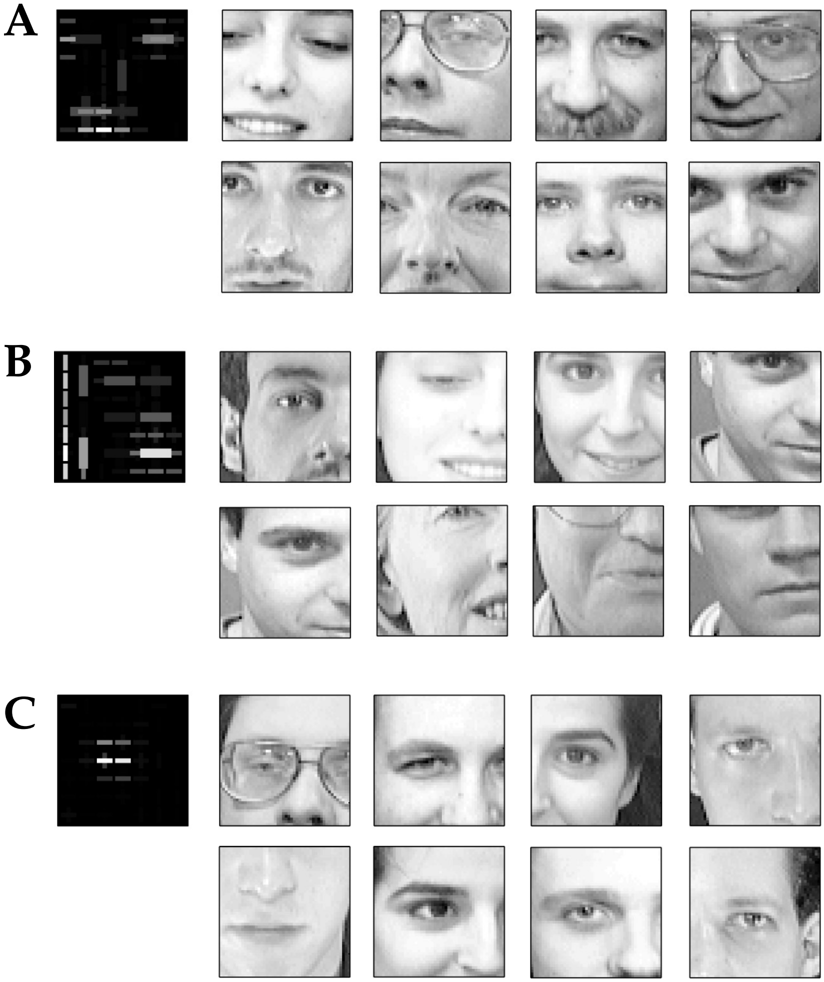


Figure 17: Example mixer association probabilities and maximal patches for face images (Samaria & Harter, 1994) and wavelets.

ideas (e.g., GSM) are converging. The resulting new insights inform a wealth of models and concepts and form the essential backdrop for the work in this article. They also link to engineering results in image coding and processing.

Our approach to the hierarchical representation of natural images was motivated by two critical factors. First, we sought to use top-down models to understand bottom-up hierarchical structure. As compellingly argued by Wainwright and Simoncelli (2000; Wainwright et al., 2001), Portilla et al. (2001, 2003), and Hyvarinen et al. (2003) in their bubbles framework, the popular GSM model is suitable for this because of the transparent statistical interplay of its components. This is perhaps by contrast with other powerful generative statistical approaches such as that of De Bonet and Viola (1997). Second, as also in Karklin and Lewicki, we wanted to learn the pattern of the hierarchical structure in an unsupervised manner. We suggested a novel extension to the GSM generative model in which mixer variables (at one level of the hierarchy) enjoy probabilistic assignments to mixture inputs (at a lower level). We showed how these assignments can be learned using Gibbs sampling. Williams and Adams (1999) used Gibbs sampling for solving a related assignment problem between child and parent nodes in a dynamical tree. Interestingly, Gibbs sampling has also been used for inferring the individual linear filters of a wavelet structure, assuming a sparse prior composed of a mixture of gaussian and Dirac delta function (Sallee & Olshausen, 2003), but not for resolving mixture associations.

We illustrated some of the attractive properties of the technique using both synthetic data and natural images. Applied to synthetic data, the technique found the proper association probabilities between the filter responses and mixer variables, and the statistics of the two GSM components (mixer and gaussian) matched the actual statistics that generated the data (see Figures 7 and 8). Applied to image data, the resulting mixer association neighborhoods showed phase invariance like complex cells in the visual cortex and showed a rich behavior of grouping along other features (that depended on the image class). The statistics of the inferred GSM components were a reasonable match to the assumptions embodied in the generative model. These two components have previously been linked to possible neural correlates. Specifically, the gaussian variable of the GSM has characteristics resembling those of the output of divisively normalized simple cells (Schwartz & Simoncelli, 2001); the mixer variable is more obviously related to the output of quadrature pair neurons (such as orientation energy or motion energy cells, which may also be divisively normalized). How these different information sources may subsequently be used is of great interest. Some aspects of our results are at present more difficult to link strongly to cortical physiology, such as the local contour versus more global patterns of orientation grouping that emerge in our and other approaches.

Of course, the model is oversimplified in a number of ways. Two particularly interesting future directions are allowing correlated filter responses and correlated mixer variables. Correlated filters are particularly important to allow overcomplete representations. Overcomplete representations have already been considered in the context of estimating a single mixer neighborhood in the GSM (Portilla et al., 2003) and in recent energy-based models (Osindero, Welling, & Hinton, 2006). They are fertile ground for future investigation within our framework of multiple mixers. Correlations among the mixer variables could extend and enrich the statistical structure in our model and are the key route to further layers in the hierarchy. As a first stage, we might consider a single extra layer that models a mixing of the mixers, prior to mixing the mixer and gaussian variables.

In our model, the mixer variables themselves are uncorrelated, and dependencies arise through discrete mixer assignments. Just as in standard statistical modeling, some dependencies are probably best captured with discrete mixtures and others with continuous ones. In this regard, it is interesting to compare our method to the strategy adopted by Karklin and Lewicki (2003a). Rather than having binary assignments arising from a mixture model, they accounted for the dependence in the filter responses by deriving the (logarithms of the) values of all the mixers for a particular patch from a smaller number of underlying random variables that were themselves mixed using a set of basis vectors. Our association probabilities reveal hierarchical structure in the same way that their basis vectors do, and indeed there are some similarities in the higher-level structures that result. For example, Karklin and Lewicki obtain either global spatial grouping favoring roughly one orientation or spatial frequency range or local spatial grouping at all orientations and frequencies. We also obtain similar results for the generic image ensembles, but our spatial groupings sometimes show orientation preference.

The relationship between our model and Karklin and Lewicki's is similar to that between the discrete mixture of experts model of Jacobs, Jordan, Nowlan, and Hinton (1991) and the continuous model of Jacobs, Jordan, and Barto (1991). One characteristic difference between these models is that the discrete versions (like ours) are more strongly competitive, with the mixer associated with a given group having to explain all their variance terms by itself. The discrete nature of mixer assignments in our model also led to a simple implementation of a powerful inference tool.

There are also other directions to pursue. First, various interesting bottom-up approaches to hierarchical representation are based on the idea that higher-level structure changes more slowly than low-level structure (Földiak, 1991; Wallis & Baddeley, 1997; Becker, 1999; Laurenz & Sejnowski, 2002; Einhäuser, Kayser, König, & Körding, 2002; Körding, Kayser, Einhäuser, & König, 2003). Although our results (and others like them; Hyvärinen & Hoyer, 2000b) show that temporal coherence is not necessary for extracting features like phase invariance, it would certainly be interesting to capture correlations between mixer variables over time as well as over space. Understanding recurrent connections within cortical areas, as studied in a bottom-up framework by Li (2002), is also key work for the future.

Second, as in applications in computer vision, inference at higher levels of a hierarchical model can be used to improve estimates at lower levels, for instance, removing noise. It would be interesting to explore combined bottom-up and top-down inference as a model for combined feedforward and feedback processing in the cortex. It is possible that a form of predictive coding architecture could be constructed, as in various previous suggestions (MacKay, 1956; Srinivasan, Laughlin, & Dubs, 1982; Rao & Ballard, 1999), in which only information not known to upper levels of the hierarchy would be propagated. However, note the important distinction between the generative model and recognition processes, such as predictive coding, that perform inference with respect to the generative model. In this article, we focused on the former.

We should also mention that not all bottom-up approaches to hierarchical structure fit into the GSM framework. In particular, methods based on discriminative ideas such as the Neocognitron (Fukushima, 1980) or the MAX model (Riesenhuber & Poggio, 1999) are hard to integrate directly within the scheme. However, some basic characteristics of such schemes, notably the idea of the invariance of responses at higher levels of the hierarchy, are captured in our hierarchical generative framework.

Finally, particularly since there is a wide spectrum of hierarchical models, all of which produce somewhat similar higher-level structures, validation remains a critical concern. Understanding and visualizing high-level, nonlinear, receptive fields is almost as difficult in a hierarchical model as it is for cells in higher cortical areas. The advantages for the model—that one can collect as many data as necessary and that the receptive fields arise from a determinate computational goal—turn out not to be as useful as one might like. One validation methodology, which we have followed here, is to test the statistical model assumptions in relation to the statistical properties of the inferred components of the model. We have also adopted the maximal patch demonstration of Karklin and Lewicki (2003a), but the results are inevitably qualitative. Other known metrics in the image processing literature, which would be interesting to explore in future work, include denoising, synthesis, and segmentation.

Appendix: Gibbs Sampling

We seek the association probabilities p_{ij} between filter response (*ie* mixture) l_i and mixer v_j that maximize the log likelihood

$$\langle \log p[\mathbf{l}|\{p_{ij}\}]\rangle_{\mathbf{l}} \tag{A.1}$$

averaged across all the input cases $\mathbf{l}$. As in the expectation maximization algorithm (Dempster et al., 1977), this involves an inner loop (the E phase), calculating the distribution of the (binary) assignments $\chi = \{\chi_{ij}\}$ for each given input patch $P[\chi|\mathbf{l}]$, and an outer loop (a partial M phase), which in this case sets new values for the association probability p_{ij} closer to the empirical mean over the E step:

$$\langle P[\chi_{ij}|\mathbf{l}]\rangle_{\mathbf{l}}. \tag{A.2}$$

We use Gibbs sampling for the E phase. This uses a Markov chain to generate samples of the binary assignments $\chi \sim P[\chi|\mathbf{l}]$ for a given input. In any given assignment, define $\eta_j = \sum_i \chi_{ij}$ to be the number of filters assigned to mixer j and $\lambda_j = \sqrt{\sum_i \chi_{ij} l_i^2}$ to be the square root of the power assigned to mixer j. Then, by the same integrals that lead to the posterior probabilities in section 3,

$$\log p[\mathbf{l}|\chi] = \log \int p[\mathbf{l}, v|\chi]dv = \log \int \prod_j p[v_j]p[\mathbf{l}|v, \chi]dv. \tag{A.3}$$

For Rayleigh prior $a = 1$, we have

$$\log p[\mathbf{l}|\chi] = \mathcal{K} + \sum_j (1 - \eta_j/2) \log \lambda_j + \log \mathcal{B}((1 - \eta_j/2), \lambda_j), \tag{A.4}$$

where $\mathcal{K}$ is a constant.

For the Gibbs sampling, we consider one filter i^* at random, and, fixing all the other assigments, $\chi^{\bar{i}^*} = \left\{\chi_{ij}, \forall i \neq i^*\right\}$, we generate a new assignment χ' according to the probabilities

$$P\left[\chi'_{i^*j} = 1, \chi^{\bar{i}^*}\right] \propto p\left[\mathbf{l}|\left\{\chi_{i^*j} = 1, \chi^{\bar{i}^*}\right\}\right]. \tag{A.5}$$

We do this many times to try to get near to equilibrium for this Markov chain and then can generate sample assignments that approximately come from the distribution $P[\chi|\mathbf{l}]$. We then use these to update the association probabilities

$$p'_{ij} = p_{ij} + \gamma \left(\langle P[\chi_{ij}|\mathbf{l}]\rangle_{\mathbf{l}} - p_{ij}\right), \tag{A.6}$$

using only a partial M step because of the approximate E step.

Acknowledgments

This work was funded by the HHMI (O.S., T.J.S.) and the Gatsby Charitable Foundation (P.D.). We are very grateful to Patrik Hoyer, Mike Lewicki, Zhaoping Li, Simon Osindero, Javier Portilla, and Eero Simoncelli for discussion.

References

Adams, N. J., & Williams, C. K. I. (2003). Dynamic trees for image modelling. *Image and Vision Computing, 10*, 865–877.

Andrews, D., & Mallows, C. (1974). Scale mixtures of normal distributions. *J. Royal Stat. Soc., 36*, 99–102.

Attneave, F. (1954). Some informational aspects of visual perception. *Psych. Rev., 61*, 183–193.

Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. In W. A. Rosenblith (Ed.), *Sensory communication.* Cambridge, MA: MIT Press.

Becker, S. (1999). Implicit learning in 3D object recognition: The importance of temporal context. *Neural Computation, 11*(2), 347–374.

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

Bollerslev, T., Engle, K., & Nelson, D. (1994). ARCH models. In B. Engle & D. McFadden (Eds.), *Handbook of econometrics V*. Amsterdam: North-Holland.

Brehm, H., & Stammler, W. (1987). Description and generation of spherically invariant speech-model signals. *Signal Processing, 12*, 119–141.

Buccigrossi, R. W., & Simoncelli, E. P. (1999). Image compression via joint statistical characterization in the wavelet domain. *IEEE Trans. Image Proc., 8*(12), 1688–1701.

De Bonet, J., & Viola, P. (1997). A non-parametric multi-scale statistical model for natural images. In M. Mozer, M. Jordan, & T. Petsche (Eds.), *Advances in neural information processing systems, 9*. Cambridge, MA: MIT Press.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood for incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39*(1), 1–38.

Einhäuser, W., Kayser, C., König, P., & Körding, K. P. (2002). Learning the invariance properties of complex cells from natural stimuli. *Eur. J. Neurosci., 15*(3), 475–486.

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Am. A, 4*(12), 2379–2394.

Field, D. J. (1994). What is the goal of sensory coding? *Neural Computation, 6*, 559–601.

Földiak, P. (1991). Learning invariance from transformation sequences. *Neural Computation, 3*, 194–200.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern., 36*, 193–202.

Geisler, W. S., Perry, J. S., Super, B. J., & Gallogly, D. P. (2001). Edge co-occurrence in natural images predicts contour grouping performance. *Vision Research, 41*, 711–724.

Grenander, U., & Srivastava, A. (2002). Probability models for clutter in natural images. *IEEE Trans. on Patt. Anal. and Mach. Intel., 23*, 423–429.

Hinton, G. E., & Ghahramani, Z. (1997). Generative models for discovering sparse distributed representations. *Philosophical Transactions Royal Society B, 352*, 1177–1190.

Hinton, G. E., Ghahramani, Z., & Teh, Y. W. (1999). Learning to parse images. In M. S. Kearns, S. A. Solla, & D. A. Cohn (Eds.), *Advances in neural information processing systems, 11* (pp. 463–469). Cambridge, MA: MIT Press.

Hoyer, P., & Hyvärinen, A. (2002). A multi-layer sparse coding network learns contour coding from natural images. *Vision Research, 42*(12), 1593–1605.

Huang, J., & Mumford, D. (1999). Statistics of natural images and models. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (p. 547). Fort Collins, CO: Computer Science Press.

Hubel, D., & Wiesel, T. (1962). Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. *Journal of Physiology (London), 160*, 106–154.

Hyvärinen, A., & Hoyer, P. (2000a). Emergence of phase- and shift-invariant features by decomposition of natural images into independent subspaces. *Neural Computation, 12*, 1705–1720.

Hyvärinen, A., & Hoyer, P. (2000b). Emergence of topography and complex cell properties from natural images using extensions of ICA. In S. A. Solla, T. K. Leen, & K.-R. Müller (Eds.), *Advances in neural information processing systems, 12* (pp. 827–833). Cambridge, MA: MIT Press.

Hyvärinen, A., Hurri, J., & Vayrynen, J. (2003). Bubbles: A unifying framework for low-level statistical properties of natural image sequences. *Journal of the Optical Society of America A, 20*, 1237–1252.

Jacobs, R. A., Jordan, M. I., & Barto, A. G. (1991). Task decomposition through competition in a modular connectionist architecture: The what and where vision tasks. *Cognitive Science, 15*, 219–250.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation, 3*, 79–87.

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

Karklin, Y., & Lewicki, M. S. (2003b). A Model for learning variance components of natural images. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems, 15* (pp. 1367–1374). Cambridge, MA: MIT Press.

Karklin, Y., & Lewicki, M. S. (2005). A hierarchical Bayesian model for learning nonlinear statistical regularities in nonstationary natural signals. *Neural Computation, 17*, 397–423.

Körding, K. P., Kayser, C., Einhäuser, W., & König, P. (2003). How are complex cell properties adapted to the statistics of natural scenes? *Journal of Neurophysiology, 91*(1), 206–212.

Laurenz, W., & Sejnowski, T. (2002). Slow feature analysis: Unsupervised learning of invariances. *Neural Computation, 14*(4), 715–770.

Lee, J. S. (1980). Digital image enhancement and noise filtering by use of local statistics. *IEEE Pat. Anal. Mach. Intell. PAMI-2*, 165–168.

Li, Z. (2002). A saliency map in primary visual cortex. *Trends in Cognitive Sciences, 6*, 9–16.

Li, Z., & Atick, J. J. (1994). Towards a theory of striate cortex. *Neural Computation, 6*, 127–146.

MacKay, D. (2003). *Information theory, inference, and learning algorithms*. Cambridge: Cambridge University Press.

MacKay, D. M. (1956). In C. E. Shannon & I. McCarthy (Eds.), *Automata studies* (pp. 235–251). Princeton, NJ: Princeton University Press.

Nadal, J. P., & Parga, N. (1997). Redundancy reduction and independent component analysis: Conditions on cumulants and adaptive approaches. *Neural Computation, 9*, 1421–1456.

Neisser, U. (1967). *Cognitive psychology*. Englewood Cliffs, NJ: Prentice Hall.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse factorial code. *Nature, 381*, 607–609.

Osindero, S., Welling, M., & Hinton, G. E. (2006). Topographic product models applied to natural scene statistics. *Neural Computation, 18*(2), 381–414.

Park, H. J., & Lee, T. W. (2005). Modeling nonlinear dependencies in natural images using mixture of Laplacian distribution. In L. K. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in neural information processing systems, 17* (pp. 1041–1048). Cambridge, MA: MIT Press.

Portilla, J., & Simoncelli, E. P. (2003). Image restoration using gaussian scale mixtures in the wavelet domain. In *Proc. 10th IEEE Int'l. Conf. on Image Proc.* (Vol. 2, pp. 965–968). Piscataway, NJ: IEEE Computer Society.

Portilla, J., Strela, V., Wainwright, M., & Simoncelli, E. (2001). Adaptive Wiener Denoising using a gaussian scale mixture model in the wavelet domain. In *Proc. 8th IEEE Int'l. Conf. on Image Proc.* (pp. 37–40). Piscataway, NJ: IEEE Computer Society.

Portilla, J., Strela, V., Wainwright, M., & Simoncelli, E. P. (2003). Image denoising using a scale mixture of gaussians in the wavelet domain. *IEEE Trans. Image Processing, 12*(11), 1338–1351.

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience, 2*(1), 79–87.

Rao, R. P. N., Olshausen, B. O., & Lewicki, M. S. (2002). *Probabilistic models of the brain*. Cambridge, MA: MIT Press.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience, 2*, 1019–1025.

Romberg, J., Choi, H., & Baraniuk, R. (1999). Bayesian wavelet domain image modeling using hidden Markov trees. In *Proc. IEEE Int'l. Conf. on Image Proc.* Piscataway, NJ: IEEE Computer Society.

Romberg, J., Choi, H., & Baraniuk, R. (2001). Bayesian tree-structured image modeling using Wavelet-domain hidden Markov models. *IEEE Trans. Image Proc., 10*(7), 1056–1068.

Ruderman, D. L., & Bialek, W. (1994). Statistics of natural images: Scaling in the woods. *Phys. Rev. Letters, 73*(6), 814–817.

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

Sallee, P., & Olshausen, B. A. (2003). Learning sparse multiscale image representations. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems 15* (pp. 1327–1334). Cambridge, MA: MIT Press.

Samaria, F., & Harter, A. (1994). Parameterisation of a stochastic model for human face identification. In *Proc. of 2nd IEEE Workshop on Applications of Computer Vision*. Piscataway, NJ: IEEE.

Schwartz, O., Sejnowski, T. J., & Dayan, P. (2005). Assignment of multiplicative mixtures in natural images. In L. K. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in neural information processing systems, 17* (pp. 1217–1224). Cambridge, MA: MIT Press.

Schwartz, O., & Simoncelli, E. P. (2001). Natural signal statistics and sensory gain control. *Nature Neuroscience, 4*(8), 819–825.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*, 379–423.

Shannon, C. E., & Weaver, W. (1949). *Nonlinear problems in random theory*. Urbana: University of Illinois Press.

Simoncelli, E. P. (1997). Statistical models for images: Compression, restoration and synthesis. In *Proc. 31st Asilomar Conf. on Signals, Systems and Computers* (pp. 673–678). Pacific Grove, CA: IEEE Computer Society. Available online at http://www.cns.nyu.edu/∼eero/publications.html

Simoncelli, E. P., Freeman, W. T., Adelson, E. H., & Heeger, D. J. (1992). Shiftable multi-scale transforms. *IEEE Trans. Information Theory, 38*(2), 587–607.

Simoncelli, E., & Olshausen, B. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience, 24*, 1193–1216.

Srinivasan, M. V., Laughlin, S. B., & Dubs, A. (1982). Predictive coding: A fresh view of inhibition in the retina. *J. R. Soc. Lond. B, 216*, 427–459.

Strela, V., Portilla, J., & Simoncelli, E. (2000). Image denoising using a local gaussian scale mixture model in the wavelet domain. In A. Aldroubi, A. F. Laine, & M. A. Unser (Eds.), *Proc. SPIE, 45th Annual Meeting*. Bellingham, WA: International Society for Optional Engineering.

Touryan, J., Lau, B., & Dan, Y. (2002). Isolation of relevant visual features from random stimuli for cortical complex cells. *J. Neurosci., 22*(24), 10811–10818.

Turiel, A., Mato, G., Parga, N., & Nadal, J. P. (1998). The self-similarity properties of natural images resemble those of turbulent flows. *Phys. Rev. Lett., 80*, 1098–1101.

van Hateren, J. H., & van der Schaaf, A. (1998). Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc. R. Soc. Lond. B, 265*, 359–366.

Wainwright, M. J., & Simoncelli, E. P. (2000). Scale mixtures of gaussians and the statistics of natural images. In S. A. Solla, T. K. Leen, & K.-R. Müller (Eds.), *Advances in neural information processing systems, 12* (pp. 855–861). Cambridge, MA: MIT Press.

Wainwright, M. J., Simoncelli, E. P., & Willsky, A. S. (2001). Random cascades on wavelet trees and their use in modeling and analyzing natural imagery. *Applied and Computational Harmonic Analysis, 11*(1), 89–123.

Wallis, G., & Baddeley, R. J. (1997). Optimal, unsupervised learning in invariant object recognition. *Neural Computation, 9*, 883–894.

Wegmann, B., & Zetzsche, C. (1990). Statistical dependence between orientation filter outputs used in an human vision based image code. In M. Kunt (Ed.), *Proc. SPIE Visual Comm. and Image Processing* (Vol. 1360, pp. 909–922). Bellingham, WA: SPIE.

Williams, C. K. I., & Adams, N. J. (1999). Dynamic trees. In M. S. Kearns, S. A. Solla, & D. A. Cohn (Eds.), *Advances in neural information processing systems, 11* (pp. 634–640). Cambridge, MA: MIT Press.

Zetzsche, C., & Nuding, U. (2005). Nonlinear and higher-order approaches to the encoding of natural scenes. *Network: Computation in Neural Systems, 16*(2–3), 191–221.

Zetzsche, C., Wegmann, B., & Barth, E. (1993). Nonlinear aspects of primary vision: Entropy reduction beyond decorrelation. In J. Morreale (Ed.), *Int'l. Symposium, Society for Information Display* (Vol. 24, pp. 933–936). Playa del Ray, CA.

Received November 18, 2004; accepted April 25, 2006.