Statistical Models of Natural Images and Cortical Visual Representation

Aapo Hyvärinen

*Department of Mathematics and Statistics, Department of Computer Science and HIIT,
Department of Psychology, University of Helsinki*

Abstract

A fundamental question in visual neuroscience is: Why are the response properties of visual neurons as they are? A modern approach to this problem emphasizes the importance of adaptation to ecologically valid input, and it proceeds by modeling statistical regularities in ecologically valid visual input (natural images). A seminal model was linear sparse coding, which is equivalent to independent component analysis (ICA), and provided a very good description of the receptive fields of simple cells. Further models based on modeling residual dependencies of the ''independent'' components have later been introduced. These models lead to emergence of further properties of visual neurons: the complex cell receptive fields, the spatial organization of the cells, and some surround suppression and Gestalt effects. So far, these models have concentrated on the response properties of neurons, but they hold great potential to model various forms of inference and learning.

Keywords: Natural image statistics; Natural scenes; Computational models; Probabilistic models; Vision

1. Introduction

Modeling images or image patches using statistical generative models has recently emerged as an established area of research (Hyvärinen, Hurri, & Hoyer, 2009; Olshausen, 2003; Simoncelli & Olshausen, 2001). This is based on the widespread assumption that biological visual systems are adapted to process the particular kind of information they receive (Barlow, 1961; Field, 1994). Natural images have important statistical regularities

Correspondence should be sent to Aapo Hyvärinen, Department of Mathematics and Statistics, Department of Computer Science and HIIT, Department of Psychology, University of Helsinki, Helsinki, Finland. E-mail: aapo.hyvarinen@helsinki.fi

that distinguish them from other kinds of input. For example, the grayscale values or luminances at different pixels have robust and nontrivial statistical dependencies. Models of the natural image statistics show what a statistically adapted representation of visual input is, indicating what the visual system should be like if it followed the assumption of optimal adaptation. Optimality is here defined in the sense of low-level signal processing operations (e.g., denoising), and it does not take into account the behavioral relevance of the signal.

Statistical models of natural images thus enable us to provide (one) answer to the fundamental question: *Why are the response properties of visual neurons as they are?* Previous theories, such as edge detection and space-frequency analysis, are unsatisfactory because they only give vague qualitative predictions on how the visual neurons should respond to visual stimulation. Statistical models offer exact quantitative prediction that often turn out to be very much in line with measurements from the visual cortex.

In the following, I provide a review of some basic models of natural images, and I point out two obvious topics that have not yet received a lot of attention: inference, and continuous updating of the representation.
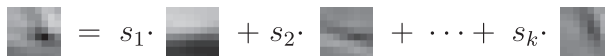
2. Linear models of natural images

The starting point in this modeling endeavor is a generative model of natural images, that is, ecologically valid visual input. Denote by $I(x,y)$ the pixel grayscale values (point luminances) in an image, or in practice, a small image patch. The models that we consider here express each image patch as a linear superposition of some features or basis vectors A_i:

$$I(x, y) = \sum_{i=1}^{n} A_i(x, y)s_i \tag{1}$$

for all x and y. The s_i are stochastic coefficients, different from patch to patch; they typically model simple-cell responses (outputs). The linear representation is illustrated in Fig. 1.

For simplicity, it is often assumed that the number of pixels equals the number of features, in which case the linear system in Eq. 1 can be inverted (but see Olshausen & Field, 1997). Then, each s_i can be computed from the image by applying a linear transformation obtained by inverting the system given by the A_i. Denoting by W_i the obtained coefficients, we have

Fig. 1. An illustration of the linear superposition in Eq. 1. On the left-hand side of this equation, we have an image patch *I*. It is represented as a weighted sum of features A_i, which are themselves plotted as image patches on the right-hand side of this equation. The weights in the summing are given by the s_i, which model simple-cell responses. The features are summed pixel by pixel to produce the image patch *I*.

$$s_i = \sum_{x,y} W_i(x,y) I(x,y). \tag{2}$$

The coefficients W_i can be interpreted as a model of the receptive field of the simple cell with index i. It can be shown (Hyvärinen & Hoyer, 2001) that the A_i are basically low-pass filtered versions of the receptive fields W_i. Therefore, most properties of the W_i and A_i are identical.

Estimation of the model consists of determining the values of A_i, observing a sufficient number of natural image patches textitI (but without observing the responses s_i). This is equivalent to determining the values of W_i, or the values of s_i for each image patch.

Such an estimation requires that we specify the statistical properties of the s_i. If we completely specify their statistical distribution, we have a well-defined statistical model and established statistical theory can be used to estimate the parameters A_i. Alternatively, in some cases we specify the statistical distribution of the s_i only to some extent, and find the A_i by maximizing some statistical property of the s_i. Sparseness, which we discuss next, is one example of such a simplification.

3. Sparseness

A considerable proportion of the models on natural image statistics is based on one particular statistical property, sparseness, which is closely related to the properties of super-Gaussianity or leptokurtosis (Field, 1994), and to independent component analysis (ICA) discussed below. The outputs of linear filters that model simple-cell receptive fields are very sparse; in fact, they are the receptive fields that maximize a suitable defined measure of sparseness.

Sparseness means that a random variable takes very small absolute values and very large values more often than a Gaussian random variable of the same variance; to compensate, it takes values in between relatively more rarely. (We assume here and in what follows that the variable has zero mean.) Thus, the random variable is "activated," that is, significantly non-zero, only rarely. This is illustrated in Fig. 2.

The probability density function (histogram) of a sparse variable is characterized by a large value (peak) at zero and relatively large values far from zero (heavy tails); see Fig. 3. Here, "relatively" means compared with a Gaussian distribution of the same variance.

Sparseness is not dependent on the variance (scale) of the random variable. To measure the sparseness of a random variable s with zero mean, let us first normalize its scale so that the variance $E\{s^2\}$ equals some given constant. Then the sparseness can be measured as the expectation $E\{G(s^2)\}$ of a suitable nonlinear function of the square. Typically, G is chosen to be convex, which means that the graph of G has a convex shape (its derivative is increasing). Convexity implies that this expectation is large when s^2 typically takes values that are either very close to 0 or very large, that is, when s is sparse (Hyvärinen et al., 2009).

For example, if G is the square function (a typical convex function), sparseness is measured by the fourth moment $E\{s^4\}$. This is closely related to using the classical fourth-order
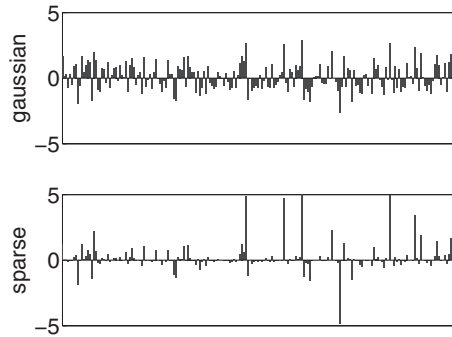
Fig. 2. Illustration of sparseness. Random samples of a Gaussian variable (top) and a sparse variable (bottom). The sparse variable is practically zero most of the time, occasionally taking very large values. Note that the variables have the same variance, and that the time structure is irrelevant in the definition of sparseness.

cumulant called kurtosis, defined as $\text{kurt}(s){=}E\{s^4\}{-}3(E\{s^2\})^2$. However, kurtosis suffers from some adverse statistical properties (Hyvärinen, Karhunen, & Oja, 2001), which is why in practice other functions may have to be used. Estimation-theoretic considerations show that in some ways the ideal function would be such that $G(s^2)$ is equal to the logarithm of a sparse probability density function, optimally of s itself. A widely used choice is $G(s^2) = \log \cosh \sqrt{s^2} = \log \cosh s$.

4. Independent component analysis

Maximization of sparseness is, in fact, very closely related to estimation of the model called independent component analysis (Comon, 1994; Hyvärinen et al., 2001). The central point in ICA is the interpretation of the linear mixing in Eq. 1 as a statistical generative
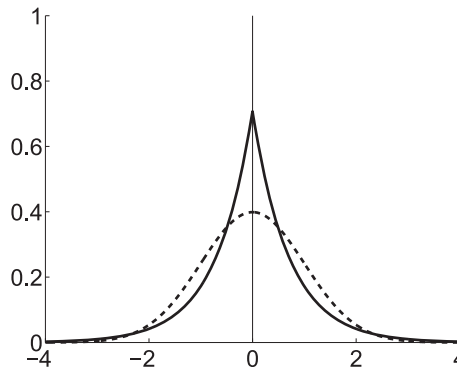

Fig. 3. Illustration of a sparse probability density. Vertical axis: probability density. Horizontal axis: (absolute) value of random variable s. The sparse density function, which has the Laplacian distribution (Hyvärinen et al., 2001), is given by the solid curve. For comparison, the density of the absolute value of a Gaussian random variable of the same variance is given by the dash-dotted curve.

model, that is, we assume that the image data I is really generated according to such a linear superposition. The vectors A_i are considered parameters in this statistical model, and the s_i are latent random variables. Then, once we define what the joint probability distribution of the s_i is, we have defined a statistical model of the observed data.

As its name says, the key idea is to define the distribution of the s_i by assuming that the s_i are mutually statistically independent. A fundamental theorem in the theory of ICA says that if the components s_i are not only independent, but also *non-Gaussian*, then the model can actually be estimated (Comon, 1994). In other words, we can recover the vectors A_i if the data were actually generated by the model. Note that the model cannot be estimated for Gaussian data.

It is important to understand the difference between independence and uncorrelatedness. Two random variables y_1 and y_2 are independent if information on the value of y_1 does not give any information on the value of y_2, and vice versa. If the two random variables are independent, they are necessarily uncorrelated as well. However, it is quite possible to have random variables that are uncorrelated, yet strongly dependent. This is illustrated in Fig. 4. In fact, if two random variables y_i and y_j were independent, any *nonlinear* transformations of them would be uncorrelated as well, which clearly shows that independence is a stronger property.

As a first step towards independence, many ICA algorithms constrain the estimated components to be exactly uncorrelated, which is a computationally simple operation. However, there is always an infinity of different representations which have uncorrelated components;
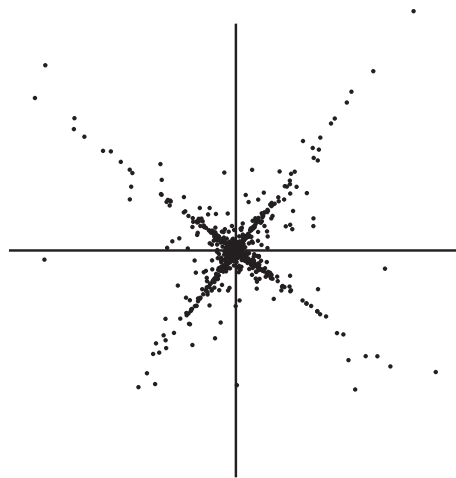

Fig. 4. Illustration of the difference between uncorrelatedness and independence. The two variables in the scatter plot are uncorrelated, but not independent. The lack of independence is due to the fact that if the variable on the horizontal axis has a large absolute value, the variable on the vertical axis has very probably a value that is far from zero. Thus, the variables provide information about each other. However, the correlation coefficient between the variables is zero because there is no *linear* trend or dependency. The dependency can be measured by taking a nonlinear transformation: in fact, the two variables have a strong correlation of their squares, which is discussed in Section 5.1.

thus, uncorrelatedness is not enough to determine the representation. In the classic literature of factor analysis, this is called the factor rotation problem.

A very deep theoretical result in ICA says that if the data are generated by the ICA model with sparse components, then the features A_i can be found by searching over all different representations and taking the one that maximizes the sparseness of the components (Comon, 1994; Hyvärinen, 1999a). Thus, sparse coding and ICA are intimately related.

The results obtained when an ICA or sparse coding model is estimated for image patches (Bell & Sejnowski, 1997; Olshausen & Field, 1996) are shown in Fig. 5. A comparison with simple-cell measurements shows quite a good match with respect to almost all parameters (van Hateren & Ruderman, 1998; van Hateren & van der Schaaf, 1998).

5. Dependencies between components

5.1. Dependency as correlation of squares

The third important statistical property used in natural image models considers the relationships between the different coefficients s_i (i.e., responses of simple cells) in Eq. 1.

In the theory of ICA and sparse coding, the s_i are usually assumed independent. However, when the models are estimated for natural images, the obtained components are *not*
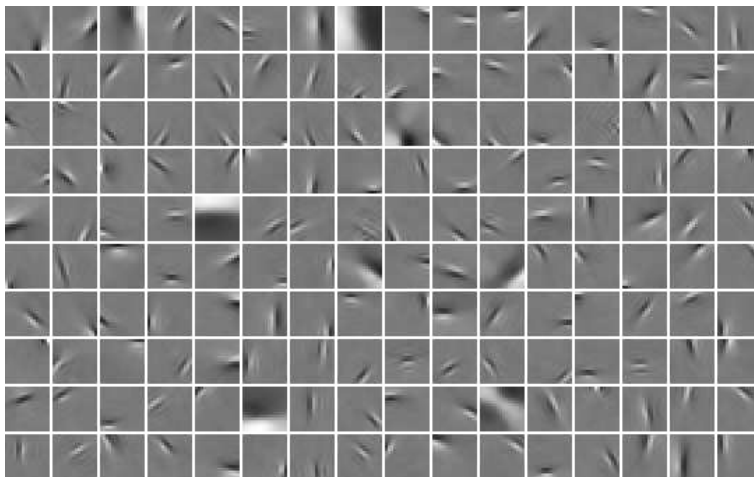

Fig. 5. Feature vectors A_i estimated by ICA/sparse coding. A set of 10,000 image patches of 16×16 pixels were randomly sampled from natural images and input to the FastICA algorithm (Hyvärinen, 1999a). Each small square gives one feature A_i, represented in the original image space so that gray pixels mean zero values (i.e., mean luminance) of $A_i(x,y)$, whereas black and white pixels mean negative and positive values, respectively. The features are shown in no particular order. See Fig. 1 for an illustrations of how these features are combined to represent an image. Before ICA estimation, the dimension of the data was reduced by principal component analysis (van Hateren & van der Schaaf, 1998; Hyvärinen et al., 2009).

independent. Basically, there are not enough parameters in the model to render the estimated linear components completely independent, as independence is a very complex phenomenon. What ICA is capable of doing is to find the linear transformation that makes the components as independent as possible by a linear transformation, but some residual dependencies still remain.

This indicates the need to model the statistical dependencies of the linear filter outputs with natural image input. Remaining dependencies actually offer a great opportunity because it means that we can hope to model further properties of visual neurons by building more sophisticated statistical models of natural images.

Note that the new models must consider *nonlinear* correlations between the components s_i. Linear correlations are not interesting in this respect because, in ICA estimation, the components are often constrained to be exactly uncorrelated. When probing the dependence of s_i and s_j, a simple approach would be to consider the correlations of some nonlinear functions. In image data, the principal form of dependency between two simple-cell responses seems to be captured by the correlation of their ''energies,'' or squares s_i^2. This means that

$$\text{cov}(s_i^2, s_j^2) = E\{s_i^2 s_j^2\} - E\{s_i^2\}E\{s_j^2\} \neq 0. \tag{3}$$

This covariance is usually positive. Intuitively, correlation of energies means that the cells tend to be active, that is, have non-zero outputs, at the same time, but the actual values of s_i and s_j are not easily predictable from each other. In Fig. 4 we already saw a scatterplot of two variables with such a dependency.

5.2. Complex cell models

The correlation of energies can be embedded in a model of natural image statistics in many ways. A very simple way would be to divide the components (simple cells) into groups, so that the s_i in the same group have correlation of energies, whereas s_i in different groups are independent.

In such a model (Hyvärinen & Hoyer, 2000), it was found that the groups (called ''independent subspaces'') show emergence of complex cells properties (see Figs. 6 and 7). The idea is to interpret the sum of squares of the s_i inside one group as a nonlinear, complex feature. After estimation of the features, these sums of squares are largely insensitive to changes in the phase of the input stimulus, that is, they give the same response to an edge feature and a bar feature, for example. Still, the sums of squares are very selective to orientation and frequency. This can be understood by noticing that the features in the same subspace have very similar orientations and frequencies (and rather similar locations), whereas their phases are quite different from each other. A complex feature (complex cell output) is thus computed by summing up responses of lower order features (simple-cell or linear filter outputs) over the dimension to which it is insensitive. It turns out that the connection to sparse coding is still there: The new model can be estimated by maximizing the sparseness of the nonlinearly pooled responses, which can be considered as models of complex cell responses.
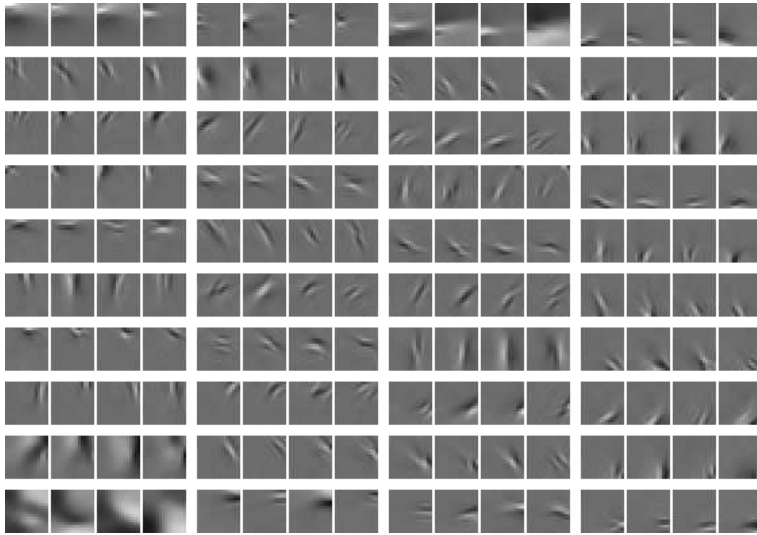
Fig. 6. Features A_i, and their grouping, estimated by independent subspace analysis. Each group of four consecutive linear features defines one nonlinear feature. It turns out that the nonlinear features have the properties of complex cells, as analyzed in Fig. 7.

The resulting responses are not unlike those in classical energy models of complex cell responses (Pollen & Ronner, 1983). Note, however, that here the model was not specified to be insensitive to phase: Natural image statistics dictated that phase is the parameter to be insensitive for.

5.3. Models of spatial organization of cells

Correlations of squares of simple-cell outputs can also be related to the spatial organization (topography) of the cells in V1 (Hyvärinen & Hoyer, 2001). Inspired by the cortical organization of cells in V1, we can arrange the components s_i on a two-dimensional grid or lattice as is typical in topographic models (Kohonen, 2001; Swindale, 1996). We can then define a statistical model in which the topographic organization reflects the statistical dependencies between the components: The components (simple cells) are arranged on the grid so that any two cells that are close to each other have dependent outputs, whereas cells that are far from each other have independent outputs. As we are using the correlation of energies as the measure of dependency, the energies are strongly positively correlated for neighboring cells.

When the model was applied on natural image data (see Fig. 8), the organization of simple cells was qualitatively very similar to the one found in the visual cortex: There is orderly arrangement with respect to such parameters as location, orientation, and spatial frequency—and no order with respect to phase (DeAngelis, Ghose, Ohzawa, & Freeman, 1999). This was the first model to show emergence of all these principal properties of cortical topography.
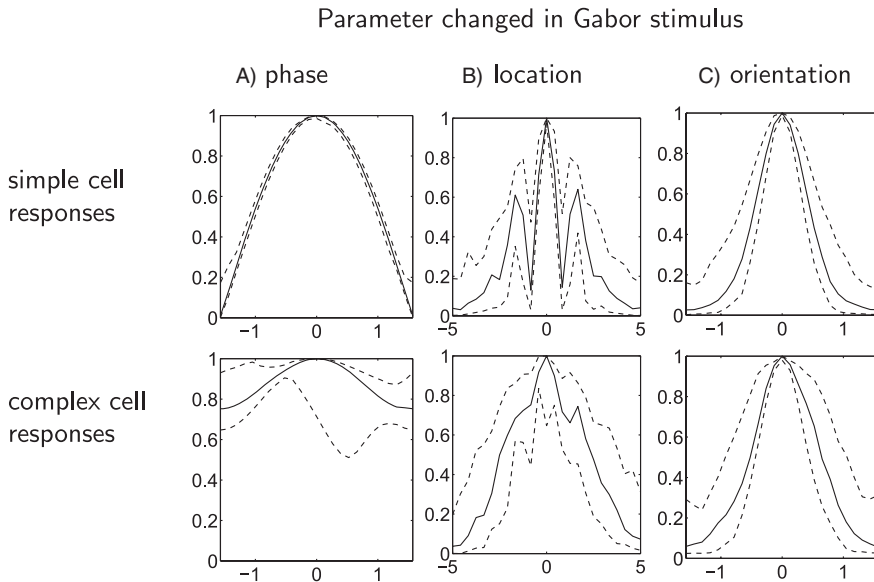
Parameter changed in Gabor stimulus

Fig. 7. Emergence of complex cell properties (Hyvärinen & Hoyer, 2000). Complex cell responses were modeled as sums of squares inside each group of four linear filters as in Fig. 6. The responses were computed for Gabor stimuli and compared with the responses of linear filters (simple cells). First, we determined the Gabor stimulus that elicited the maximum response. Then, we varied the stimulus according to one of the parameters (phase, location, or orientation). In all plots, the solid line gives the median response in the population of all cells, and the dotted lines give the 90% and 10% percentiles of the responses. Responses are normalized relative to response to preferred stimulus. Top row: responses (in absolute values) of simple cells (linear features). Bottom row: responses of complex cells (sum of squares of linear features). (A) Effect of varying phase. (B) Effect of varying location (shift). (C) Effect of varying orientation. We see that the complex cell response is insensitive to change in phase of the stimulus, while selective to the other parameters; by contrast, simple-cell responses are selective to all the parameters.

An interesting point in this model is that the topography defined by dependencies is closely related to complex cells. If we define nonlinear features by summing the squares of outputs of simple cells nearby on the topographic grid (Blasdel, 1992), we obtain, again, complex cell properties.

5.4. Further models based on correlations of squares

The dependencies of ''independent'' components can also be used to model contrast gain control phenomena. The idea is that a large part of the dependencies is due to a single factor, possibly related to lighting conditions, which influences the variances of all the linear components. Removing this factor leads to a divisive normalization, which is not very different from models of contrast gain control (Heeger, 1992). Such models can explain some surround suppression effects. They have been mainly developed in the context of wavelet-based image models (Schwartz & Simoncelli, 2001; for a recent development in the context of ICA, see Köster, Lindgren, Gutmann, & Hyvärinen, 2009).
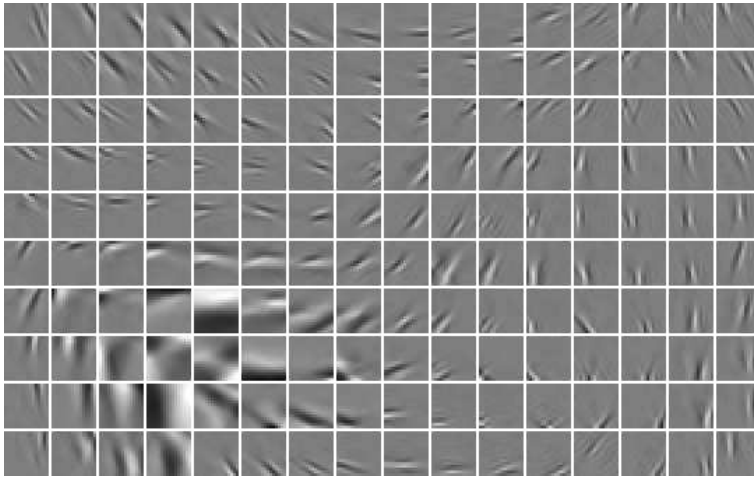
Fig. 8. Features A_i, and their topographic organization, estimated by topographic ICA.

A further line of research has linked the correlations of squares of features in different spatial locations to contour completion and Gestalt formation. In fact, collinear Gabor features, which are likely to belong to the same contour and relate to the Gestalt principle of ''good continuation,'' have strong square correlations as well (Geisler, Perry, Super, & Gallogly, 2001; Krüger, 1998; Sigman, Cecchi, Gilbert, & Magnasco, 2001; for a related review, see Geisler, 2008).

6. Different timescales of learning and inference

In the models discussed above, features were learned (estimated) for natural images. The models operate on an abstract level where, in particular, it is not specified at what point of time the learning happens. Thus, the models are completely neutral as to the nature versus nurture questions: The learning could be accomplished during evolution or during the lifetime of the organism.

Implicitly, the literature seems to assume that the representation should be fixed at least after maturation of the organism. However, this need not be so, and it is conceivable that the representation is constantly updated (relearned) so that recent input is emphasized. Online (neural) algorithms usually lead to such emphasis anyway.

Continuous updating of the representation could explain some phenomena of *perceptual learning*. Curiously, hardly any attempts have been made to model such updating based on statistics of natural images. However, the application of the models reviewed here should be rather straightforward, at least regarding learning of the representation. If we simply add a sample of the stimuli used in a perceptual learning experiment to the sample of natural image patches, the learned representation will be different.

A closely related topic is inference. In statistical models, inference usually means computing (estimating) the values of latent variables, such as the s_i in Eq. 1. In most models

discussed in this review, the transformation from the s_i to the images is invertible; thus, the values of the s_i can be immediately and exactly computed if we know the parameters A_i. However, this is not the case if, for example, there are more latent variables s_i than observed variables $I(x,y)$ (Olshausen & Field, 1997), or there is noise in the process (Hyvärinen, 1999b). In general, we can only infer the most probable s_i by computing the posterior probability $p(s_i|I,A)$.

Inference is usually assumed to happen very fast in the neural systems, as it is closely connected to computing the immediate responses of the cells to stimuli. In fact, there may be several timescales involved. Adaptation is presumably slower than inference but faster than perceptual learning, and it may involve many timescales in itself.

Inference requires that we construct a model of the distributions of the components s_i (see, e.g., Hyvärinen et al., 2009). One might assume that this model changes faster than the actual features A_i. Thus, relatively fast adaptation phenomena have been modeled using the principle of adding special data to the natural image patches as described above but updating the model of the distributions of the s_i while leaving the features A_i constant (Wainwright, Schwartz, & Simoncelli, 2002).

Finally, the dynamics at different timescales might require new models. One approach would be to consider natural image sequences instead of static images (see, e.g., Hashimoto, 2003; Hurri & Hyvärinen, 2003; Hyvärinen, Hurri, & Väyrynen, 2003). With such models, we can introduce new latent variables, which opens a new, relatively unexplored way to model the temporal dynamics on different timescales. For example, one can introduce variables that model the variances of each s_i. These variables could be time correlated. Inference of the ''original'' s_i then naturally proceeds by normalizing by these variance variables, not unlike in the case of contrast gain control (Hyvärinen et al., 2003). Such normalization may be related to adaptation (Buiatti & van Vreeswijk, 2003).

7. Discussion

Modelling the statistical structure of natural images is a modern approach to modeling receptive fields. Possibly the most fundamental model is ICA, although sparse coding is an older concept. The components obtained by them are not independent, which shows, in fact, an opportunity to model further aspects of the visual system. Thus, in addition to simple-cell receptive fields, further models lead to emergence of complex cell properties, the spatial (topographic) organization of the cells, contrast gain control, and Gestalt phenomena.

Another approach to natural images is based on temporal coherence or stability (Földiák, 1991; Hashimoto, 2003; Hurri & Hyvärinen, 2003; Wiskott & Sejnowski, 2002). This means that when the input consists of natural image *sequences*, that is, video data, the outputs of simple cells in subsequent time points should be ''coherent'' or ''stable,'' that is, change as little as possible. It is also possible to develop a unifying theoretical framework for the statistical properties discussed above: sparseness, temporal coherence, and topography; this is based on the concept of a spatiotemporal bubble (Hyvärinen et al., 2003).

A topic of intense current research is how to estimate both layers in a two-layer model of natural images. This is in stark contrast to the models discussed above that fix the pooling of simple-cell responses beforehand, and only estimate the A_i or W_i. In the models introduced above, estimation of the pooling weights is possible, in principle, by considering them as another set of parameters. However, this leads to serious technical difficulties (for recent advances, see Karklin & Lewicki, 2008; Köster & Hyvärinen, 2007; Osindero, Welling, & Hinton, 2006).

Another recently developed topic is estimation of Markov random fields, which provides a model for whole images instead of small patches (Köster, Lindgren, & Hyvärinen, 2009). The basic idea is that the same features are replicated in all possible spatial locations of the image. An alternative approach for modeling whole images is to use wavelet tranformations, thus completely fixing the features instead of learning them. The latter approach has been quite successful in engineering applications (Simoncelli, 2005; Srivastava, Lee, Simoncelli, & Chu, 2003).

It would be most useful if we could use this modeling endeavor, based on statistical models of ecologically valid stimuli, in a *predictive* manner. This means that we would be able to predict properties of cells in the visual cortex, in cases where the properties have not yet been demonstrated experimentally. Thus, we would obtain testable, quantitative hypotheses that might lead to great advances, especially in the research in extrastriate areas such as V2, whose function is not well understood at this point. One study based on such modeling (Hyvärinen, Gutmann, & Hoyer, 2005) predicted that in V2 (or some related area) there should be cells whose optimal stimulus is a broadband edge that has no side lobes while being relatively sharp, that is, the optimal stimulus is closer to a step-edge than the Gabor-like band-pass edges that tend to be optimal for V1 simple and complex cells.

References

Barlow, H. B. (1961). Possible principles underlying the transformations of sensory messages. In W. A. Rosenblith (Ed.), *Sensory communication* (pp. 217–234). Cambridge, MA: MIT Press.

Bell, A. J., & Sejnowski, T. J. (1997). The ''independent components'' of natural scenes are edge filters. *Vision Research*, *37*, 3327–3338.

Blasdel, G. G. (1992). Orientation selectivity, preference, and continuity in monkey striate cortex. *Journal of Neuroscience*, *12*(8), 3139–3161.

Buiatti, M., & van Vreeswijk, C. (2003). Variance normalisation: A key mechanism for temporal adaptation in natural vision? *Vision Research*, *43*, 1895–1906.

Comon, P. (1994). Independent component analysis—a new concept? *Signal Processing*, *36*, 287–314.

DeAngelis, G. C., Ghose, G. M., Ohzawa, I., & Freeman, R. D. (1999). Functional micro-organization of primary visual cortex: Receptive field analysis of nearby neurons. *Journal of Neuroscience*, *19*(10), 4046–4064.

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

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

Geisler, W. S. (2008). Visual perception and the statistical properties of natural scenes. *Annual Review of Psychology*, *59*, 167–192.

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.

Hashimoto, W. (2003). Quadratic forms in natural images. *Network: Computation in Neural Systems*, *14*(4), 765–788.

van Hateren, J. H., & Ruderman, D. L. (1998). Independent component analysis of natural image sequences yields spatiotemporal filters similar to simple cells in primary visual cortex. *Proceedings of the Royal Society Series B*, *265*, 2315–2320.

van Hateren, J. H., & van der Schaaf, A. (1998). Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings of the Royal Society Series B*, *265*, 359–366.

Heeger, D. (1992). Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, *9*, 181–198.

Hurri, J., & Hyvärinen, A. (2003). Simple-cell-like receptive fields maximize temporal coherence in natural video. *Neural Computation*, *15*(3), 663–691.

Hyvärinen, A. (1999a). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, *10*(3), 626–634.

Hyvärinen, A. (1999b). Sparse code shrinkage: Denoising of non-Gaussian data by maximum likelihood estimation. *Neural Computation*, *11*(7), 1739–1768.

Hyvärinen, A., Gutmann, M., & Hoyer, P. O. (2005). Statistical model of natural stimuli predicts edge-like pooling of spatial frequency channels in V2. *BMC Neuroscience*, *6*(12).

Hyvärinen, A., & Hoyer, P. O. (2000). Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces. *Neural Computation*, *12*(7), 1705–1720.

Hyvärinen, A., & Hoyer, P. O. (2001). A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images. *Vision Research*, *41*(18), 2413–2423.

Hyvärinen, A., Hurri, J., & Hoyer, P. O. (2009). *Natural image statistics*. Berlin: Springer-Verlag.

Hyvärinen, A., Hurri, J., & Väyrynen, J. (2003). Bubbles: A unifying framework for low-level statistical properties of natural image sequences. *Journal of the Optical Society of America A*, *20*(7), 1237–1252.

Hyvärinen, A., Karhunen, J., & Oja, E. (2001). *Independent component analysis*. London: Wiley Interscience.

Karklin, Y., & Lewicki, M. S. (2008). Emergence of complex cell properties by learning to generalize in natural scenes. *Nature*, *104*, 19120–19125.

Kohonen, T. (2001). *Self-organizing maps* (3rd ed.). Berlin: Springer.

Köster, U., & Hyvärinen, A. (2007). A two-layer ICA-like model estimated by score matching. In J. P. Marques de Sa, W. Duch & L. A. Alexandre (Eds.), *Proceedings of the international conference on artificial neural networks (ICANN2007)* (pp. 798–807). Porto, Portugal. Berlin: Springer-Verlag.

Köster, U., Lindgren, J. T., Gutmann, M., & Hyvärinen, A. (2009). Learning natural image structure with a horizontal product model. In T. Adali, C. Jutten, J. M. Travassos Romano, & A. Kardec Barros (Eds.), *Proceedings of the international conference on independent component analysis and blind signal separation (ICA2009)* (pp. 507–514). Paraty, Brazil, Berlin: Springer-Verlag.

Köster, U., Lindgren, J. T., & Hyvärinen, A. (2009). Estimating Markov random field potentials for natural images. In T. Adali, C. Jutten, J. M. Travassos Romano & A. Kardec Barros (Eds.), *Proceedings of the international conference on independent component analysis and blind signal separation (ICA2009)* (pp. 515–522), Paraty, Brazil, Berlin: Springer-Verlag.

Krüger, N. (1998). Collinearity and parallelism are statistically significant second order relations of complex cell responses. *Neural Processing Letters*, *8*, 117–129.

Olshausen, B. A. (2003). Principles of image representation in visual cortex. In L. M. Chalupa & J. S. Werner (Eds.), *The visual neurosciences*. Cambridge, MA: MIT Press.

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

Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, *37*, 3311–3325.

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

Pollen, D., & Ronner, S. (1983). Visual cortical neurons as localized spatial frequency filters. *IEEE Transactions on Systems, Man, and Cybernetics*, *13*, 907–916.

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

Sigman, M., Cecchi, G. A., Gilbert, C. D., & Magnasco, M. O. (2001). On a common circle: Natural scenes and gestalt rules. *Proceedings of the National Academy of Science, USA*, *98*, 1935–1940.

Simoncelli, E. P. (2005). Statistical modeling of photographic images. In A. Bovik (Ed.), *Handbook of image and video processing* (2nd ed.). London: Academic Press.

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

Srivastava, A., Lee, A. B., Simoncelli, E. P., & Chu, S.-C. (2003). On advances in statistical modelling of natural images. *Journal of Mathematical Imaging and Vision*, *18*, 17–33.

Swindale, N. V. (1996). The development of topography in the visual cortex: A review of models. *Network*, *7*(2), 161–247.

Wainwright, M. J., Schwartz, O., & Simoncelli, E. P. (2002). Natural image statistics and divisive normalization: Modeling nonlinearity and adaptation in cortical neurons. In R. P. N. Rao, B. A. Olshausen & M. S. Lewicki, (Eds.), *Probabilistic models of the brain* (pp. 203–222). Cambridge, MA: MIT Press.

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