In: M. Fahle and T. Poggio (Eds).
Perceptual Learning, MIT Press
(2002)

Learning to Find Independent Components in Natural Scenes

19

Anthony J. Bell and Terrence J. Sejnowski

Abstract

The brain may operate in an arbitrarily complex way, while its self-organizational, or learning, processes, to the extent that they can be distinguished from normal operation, may be quite simple. Therefore one research program is to define learning rules using candidate abstract principles, and apply these algorithms to natural data to see if they produce processing systems similar to those found in neurons exposed to the same data.

In this chapter, we apply information-theoretic learning to noiseless sigmoidal neurons exposed to natural images. The neurons learn localized oriented receptive fields qualitatively similar to simple cells in area V1 of visual cortex. The algorithm maximizes the information contained about the input while choosing a coordinate system that makes each element of the resulting code as independent as possible. To a first approximation, the recoding of visual input in early perceptual processing may follow simple information-theoretic principles.

Both the classic experiments of Hubel and Wiesel (1968) on neurons in visual cortex and several decades of theorizing about feature detection in vision (Marr and Hildreth 1980) have left open the question most succinctly phrased by Barlow and Tolhurst (1992) "Why do we have edge detectors?" Barlow (1989) has suggested that the line and edge selectivities of neurons found in primary visual cortex of cats and monkeys should emerge from an unsupervised learning algorithm that attempts to find a factorial code of independent visual features. Along similar lines, Field (1994) has argued that a sparse, distributed representation of natural scenes should be goal of early visual representations.

These hypotheses can now be tested with unsupervised learning algorithms whose goal is to either find maximally independent linear filters (Bell and Sejnowski 1995a, 1997) or to maximize sparseness (Olshausen and Field 1997), applied to an ensemble of natural scenes. Both of these approaches produce sets of visual filters that are localized and oriented, including some filters whose associated basis functions are Gabor-like. These results are quite different from the filters produced by other decorrelating filters produced by principal components analysis (PCA) and zero-phase components analysis (ZCA). The set of filters produced by independent component analysis (ICA) has more sparsely distributed (kurtotic) outputs on natural scenes (Bell and Sejnowski 1997). They also resemble the receptive fields of simple cells in visual cortex, which suggests that these neurons form a natural, information-theoretic coordinate system for natural images.

Most studies that have examined the statistics of natural images for the purpose of reducing redundancy, and thereby making a more efficient code, have used only the second-order statistics required for *decorrelating* the outputs of a set of feature detectors. Although Hebbian feature-learning algorithms for decorrelation have been proposed (Linsker 1992; Miller 1988; Oja 1989; Sanger 1989; Földiák 1990; Atick and Redlich 1993), in the absence of particular external constraints the solutions to the decorrelation problem are nonunique. One popular decorrelating solution is principal components analysis (PCA), but the principal components of natural scenes amount to

a global spatial frequency analysis (Hancock, Baddeley and Smith 1992). Thus, second-order statistics alone do not suffice to predict the formation of localized edge detectors.

Additional constraints are required. Field (1987, 1994) has argued for the importance of sparse, or "minimum entropy," coding (Barlow 1994), in which each feature detector is activated as rarely as possible. This has led to feature-learning algorithms with a "projection pursuit" flavor (Huber 1985, Intrator 1992, Baddeley 1996, Olshausen and Field 1997).

An alternative constraint is to start with an information-theoretic criterion that maximizes the joint entropy of a nonlinearly transformed output feature vector. This is the approach taken by "independent components analysis" (Comon 1994) which can achieve the blind separation of mixed sources (Jutten and Hérault 1991; Bell and Sejnowski 1995a, 1996). Finding independent components is equivalent to Barlow's redundancy reduction problem; therefore if Barlow's reasoning is correct, the independent components should produce filters which are localized and oriented, and in fact it does. In addition, when applied to natural images, the outputs of the resulting filters are more sparsely distributed than those of other decorrelating filters, thus supporting some of the arguments of Field (1994) and helping to explain the results of Olshausen and Field (1997) from an information-theoretic point of view.

We will return to the issues of sparseness, noise and higher-order statistics. First, we describe more concretely the filter-learning problem.

19.1 "Causes" in Natural Images

A perceptual system is exposed to a series of small image patches, drawn from one or more larger images. Imagine that each image patch, represented by the vector $\mathbf{x}$, has been formed by the linear combination of N basis functions. The basis functions form the columns of a fixed matrix, $\mathbf{A}$. The weighting of this linear combination (which varies with each image) is given by a vector, $\mathbf{s}$. Each component of this vector has its own associated basis function, and represents an underlying "cause" of the image. The *linear image synthesis* model is therefore given by:

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{19.1}$$

which is the matrix version of the set of equations

$$x_i = \sum_{j=1}^{N} a_{ij} s_j \tag{19.2}$$

where each x_i represents a pixel in an image, and contains contributions from each one of a set of N image "sources," s_j, linearly weighted by a coefficient, a_{ij}.

The goal of a perceptual system, in this simplified framework, is to linearly transform the images, $\mathbf{x}$, with a matrix of filters, $\mathbf{W}$, so that the resulting vector:

$$\mathbf{u} = \mathbf{W}\mathbf{x} \tag{19.3}$$

recovers the underlying causes, $\mathbf{s}$, possibly in a different order, and rescaled. Representing, by $\mathbf{P}$, an arbitrary permutation matrix (all zero except for a single "one" in each row and each column), and, by $\mathbf{S}$, an arbitrary scaling matrix (nonzero entries only on the diagonal), such a system has converged when:

$$\mathbf{u} = \mathbf{W}\mathbf{A}\mathbf{s} = \mathbf{P}\mathbf{S}\mathbf{s} \tag{19.4}$$

The scaling and permuting of the causes are arbitrary, unknowable factors, so consider the causes to be defined such that $\mathbf{PS} = \mathbf{I}$ (the identity matrix).

Then the basis functions (columns of $\mathbf{A}$) and the filters that recover the causes (rows of $\mathbf{W}$) have the simple relation: $\mathbf{W} = \mathbf{A}^{-1}$.

All that remains in defining an algorithm to learn $\mathbf{W}$ (and thus also $\mathbf{A}$) is to decide what constitutes a "cause." We concentrate here on algorithms producing causes that are decorrelated, and those attempting to produce causes that are statistically independent.

19.2 Decorrelation and Independence

The matrix, $\mathbf{W}$, is a *decorrelating* matrix when the covariance matrix of the output vector, $\mathbf{u}$, satisfies:

$$\langle \mathbf{u}\mathbf{u}^T \rangle = \text{diagonal matrix} \tag{19.5}$$

In general, there will be many $\mathbf{W}$ matrices which decorrelate. For example, when $\langle \mathbf{u}\mathbf{u}^T \rangle = \mathbf{I}$, then:

$$\mathbf{W}^T\mathbf{W} = \langle \mathbf{x}\mathbf{x}^T \rangle^{-1} \tag{19.6}$$

which clearly leaves freedom in the choice of $\mathbf{W}$. There are, however, several special solutions to Eq. (19.6).

Principal components analysis (PCA) is the orthogonal solution to Eq. (19.5). The principal components come from the eigenvectors of the covariance matrix. The filters are orthogonal. When the image statistics are stationary (Field 1994), the PCA filters are *global* Fourier filters, ordered according to the amplitude spectrum of the image. Example PCA filters are shown in figure 19.1a.

If $\mathbf{W}$ is forced to be symmetrical, so that $\mathbf{W}_Z^T = \mathbf{W}_Z$, then the resulting decorrelating filters are zero-phase (ZCA). ZCA is in several ways the polar opposite of PCA. It produces *local* (center-surround type) whitening filters, which are ordered according to the phase spectrum of the image. That is, each filter whitens a given pixel in the image, preserving the spatial arrangement of the image and flattening its frequency (amplitude) spectrum (Goodall 1960; Atick and Redlich 1993). Example ZCA filters and basis functions are shown in figure 19.1b.

Another way to constrain the solution is to attempt to produce outputs that are not just decorrelated but statistically independent (Jutten and Hérault 1991; Comon 1994). The values of the u_i are independent when their probability distribution, $f_{\mathbf{u}}$, factorizes: $f_{\mathbf{u}}(\mathbf{u}) = \prod_i f_{u_i}(u_i)$. There are many ICA algorithms, based on different approaches (Cardoso and Laheld 1996; Karhunen et al. 1996; Amari, Cichoki, and Yang 1996; Cichocki, Unbehauen, and Rummert 1994; Pham, Garrat, and Jutten 1992; Bell and Sejnowski 1995a).

ICA produces decorrelating filters that are sensitive to both phase (locality) and frequency information, just as in transforms involving oriented Gabor functions (Daugman 1985) or wavelets. These filters are thus semilocal, depicted in figure 19.2 as partway along the path from the local (ZCA) to the global (PCA) solutions in the space of decorrelating solutions. Example ICA filters are shown in figure 19.1d and their corresponding basis functions are shown in figure 19.1e.

It is important to recognize two differences between finding an ICA solution, $\mathbf{W}_I$, and other decorrelation methods: (1) there may be no ICA solution, and (2) a given ICA algorithm may not find the solution even if it exists, because there are approximations involved. In these senses, ICA is different from PCA and ZCA, and cannot be calculated analytically, for example, from second-order statistics (the covariance matrix), except in the Gaussian case (when second-order statistics completely characterize the signal distribution).

The approach developed in Bell and Sejnowski 1995a was to maximize by stochastic gradient ascent the joint entropy, $H[g(\mathbf{u})]$, of the linear transform

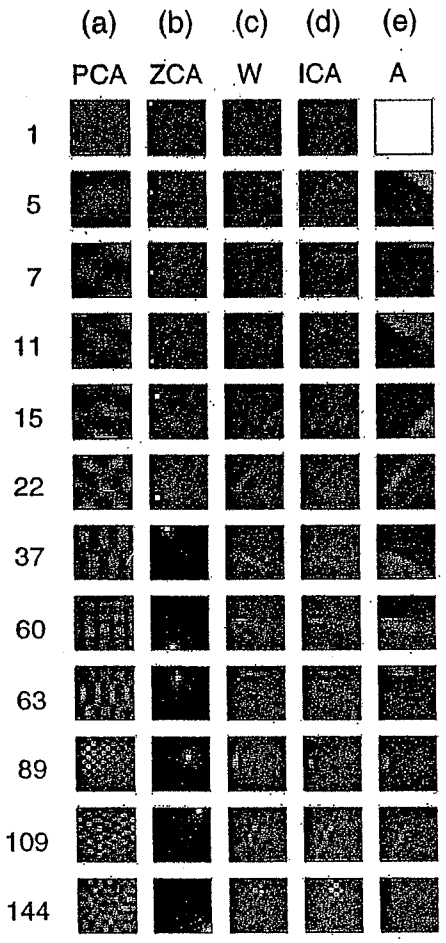

Figure 19.1
Selected decorrelating filters and their basis functions extracted from the natural scene data. Each type of decorrelating filter yielded 144 12×12 filters, of which we only display a subset here. Each column contains filters or basis functions of a particular type, and each of the rows has a number relating to which row of the filter or basis function matrix is displayed. (*a*) Principal components analysis (PCA, or $\mathbf{W}_P$): The 1st, 5th, 7th, etc. principal components, showing increasing spatial frequency. There is no need to show basis functions and filters separately here, since for PCA they are the same thing. (*b*) Zero-phase components analysis (ZCA, or $\mathbf{W}_Z$): The first six entries in this column show the one-pixel-wide center-surround filter which whitens while preserving the phase spectrum. All are identical, but shifted. The lower six entries $(37, 60 \ldots 144)$ show the basis functions instead, which are the columns of the inverse of the $\mathbf{W}_Z$ matrix. (*c*) The weights, $\mathbf{W}$, learned by the independent component analysis network trained on $\mathbf{W}_Z$-whitened data, showing (in descending order) the DC filter, localized oriented filters, and localized checkerboard filters. (*d*) The corresponding ICA filters, in the matrix $\mathbf{W}_I$, look like whitened versions of the $\mathbf{W}$-filters. (*e*) The corresponding basis functions, columns of $\mathbf{W}_I^{-1}$ (or $\mathbf{A}$). These are the patterns that optimally stimulate their corresponding ICA filters, while not stimulating any other ICA filter, so that $\mathbf{W}_I\mathbf{A} = \mathbf{I}$.

squashed by a sigmoidal function, g. When the nonlinear function is the same (up to scaling and shifting) as the cumulative density functions (CDFs) of the underlying independent components, it can be shown (Nadal and Parga 1995) that such a nonlinear "infomax" procedure also minimizes the mutual information between the u_i, exactly what is required for ICA.

However, in most cases we must pick a nonlinearity, g, without any detailed knowledge of the probability density functions (PDFs) of the underlying independent components. The resulting mismatch between the gradient of the nonlinearity used and the underlying PDFs may cause the infomax solution to deviate from an ICA solution. In cases where the PDFs are super-Gaussian (meaning they are peakier and longer-tailed than a Gaussian, having kurtosis greater than 0), we have repeatedly observed, using the logistic or hyperbolic tangent nonlinearities, that maximization of $H[g(\mathbf{u})]$ still leads to ICA solutions, when they exist, as with our experiments on speech signal separation (Bell and Sejnowski 1995a). An extended version of this algorithm can be used when there are mixed sub-Gaussian and

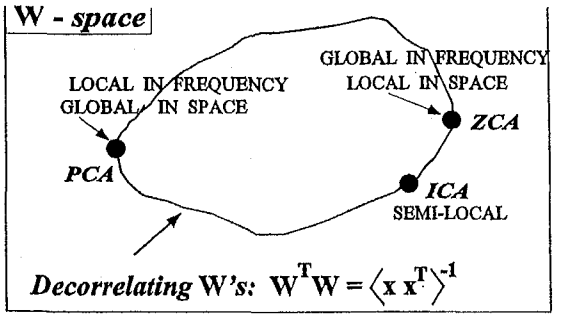

Figure 19.2
Schematic depiction of weight space. A subspace of all matrices, **W**, here represented by the loop (of course it is a much higher-dimensional closed subspace), has the property of decorrelating the input vectors, **x**. On this manifold, several special linear transformations can be distinguished: principal components analysis (PCA), global in space and local in frequency; zero-phase components analysis (ZCA), local in space and global in frequency; and independent components analysis (ICA), a privileged decorrelating matrix which, if it exists, decorrelates higher as well as second-order moments. ICA filters are localized, but not down to the single-pixel level, as ZCA filters are.

super-Gauassian sources (Lee, Girolami and Sejnowski 1999).

The filters and basis functions resulting from training on natural scenes are displayed in figures 19.1 and 19.4. Figure 19.1 displays example filters and basis functions of each type. The PCA filters (figure 19.1, panel a) are spatially global and ordered in frequency. The ZCA filters and basis functions are spatially local and ordered in phase. The ICA filters, whether trained on the ZCA-whitened images (figure 19.1, panel c) or the original images (figure 19.3, panel d) are semilocal filters, most with a specific orientation preference. The basis functions (figure 19.1, panel e), calculated from the ICA filters (figure 19.1, panel d), are not local and look like the edges that might occur in image patches of this size. Basis functions in figure 19.1, panel d, are the same as the corresponding filters because the matrix **W** is orthogonal, as is the case for the PCA filters, $\mathbf{W}_P$. This is the ICA-matrix for ZCA-whitened images.

Figure 19.4 shows, with lower resolution, all 144 filters in the matrix **W**. The general result is that ICA filters are localized and mostly oriented. There is one DC filter and fewer than ten unoriented checkerboard filters.

Figure 19.5 shows the result of analyzing the distributions (image histograms) produced by each of the three filter types. As emphasized by Ruderman (1994) and Field (1994), the general form of these histograms is double-exponential ($e^{-|u_i|}$), or "sparse," meaning peaky with a long tail, when compared to a Gaussian. This shows up clearly in figure 19.4, where the log histograms are seen to be roughly linear across twelve orders of magnitude. The histogram for the ICA filters, however, departs from linearity, having a longer tail than the ZCA and PCA histograms. This spreading of the tail signals the greater sparseness of the outputs of the ICA filters, and this is reflected in the kurtosis measure of 10.04 for ICA, compared to 3.74 for PCA, and 4.5 for ZCA.

Univariate statistics can only capture part of the story, so in figure 19.5, panels a, c and e, are displayed, in contour plots, the average of the bivariate log histograms given by all *pairs* of filters, for ICA, ZCA and PCA respectively. In contrast with these joint probability distributions, figure 19.6, panels b, d and f, show the corresponding distribution if the outputs of the filters were independent (i.e., the outer product of the marginal, or univariate, distributions in figure 19.4). Only the ICA joint histogram captures well the diamond-shape characteristic of the product of the sparse univariate distributions, thus satisfying, to a greater extent, the independence criterion: $f_{u_1 u_2}(u_1, u_2) = f_{u_1}(u_1) f_{u_2}(u_2)$.

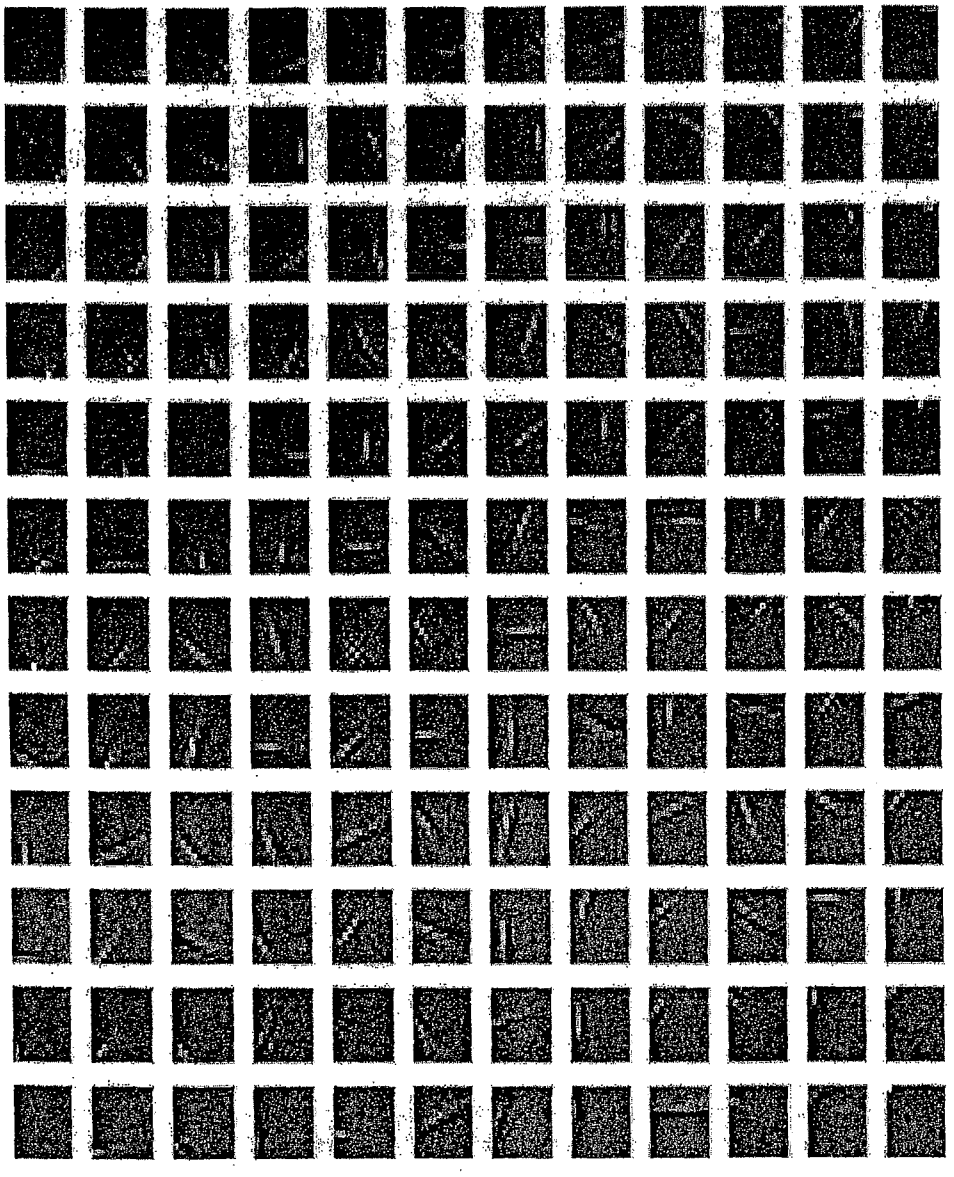

Figure 19.3
Matrix of 144 filters obtained by training on natural images whitened by zero-phase components analysis. Each filter is a row of the matrix W. The independent components analysis basis functions on ZCA-whitened data are visually the same as the ICA filters. On nonwhitened data, the filters look like high-pass versions of the filters shown here, and the basis functions look like low-pass versions of them.

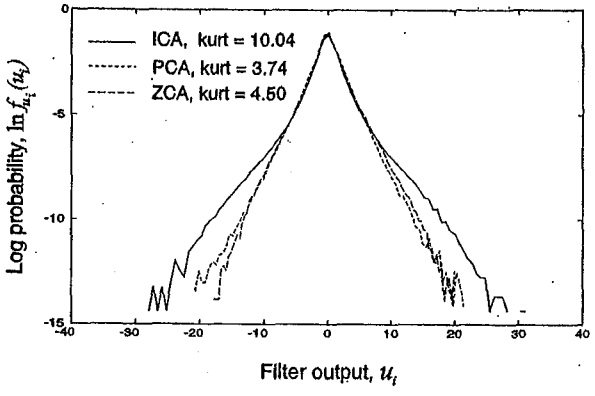

Figure 19.4
Log distributions of univariate statistics of the outputs of independent, zero-phase and principal components analysis (ICA, ZCA, and PCA) filters, averaged over all filters of each type. All three are approximately double-exponential distributions, but the more kurtotic ICA distribution is slightly peakier and has a longer tail, showing that it is *sparser* than the others. This distribution (and the two dimensional ones in figure 19.5), although averaged over the outputs of all filters, are extremely similar to the distributions output by individual filters (respectively, pairs of filters). The only exception is the DC filter (top left in 19.3) which has a more Gaussian distribution.

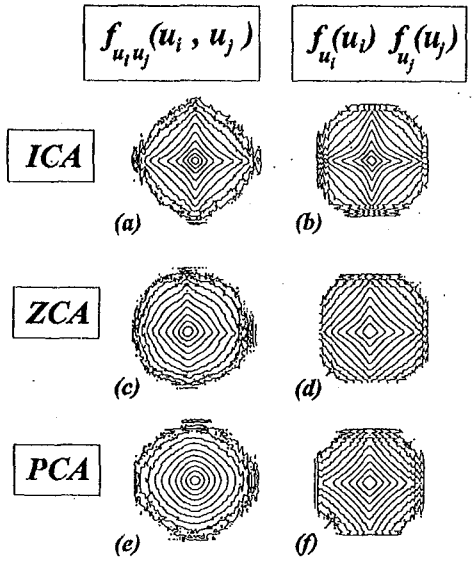

Figure 19.5
Contour plots of log distributions of pairwise statistics of the outputs of independent, zero-phase, and principal components analysis (ICA, ZCA, and PCA) filters. (*a, c, e*) Joint log distributions averaged over all pairs of output filters of each type, and all images. (*b, d, f*) Product of marginal (univariate) distributions. The ICA solution best satisfies the independence criterion that the joint distribution has the same form as the product of the marginal distributions.

In summary, the filters found by the infomax ICA algorithm with a logistic nonlinearity are localized, oriented, and produce outputs distributions of very high kurtosis.

19.3 Comparisons with Other Approaches

A substantial literature exists on the self-organization of visual receptive fields through factors such as learning. Many contributions have emphasized the roles of decorrelation and PCA (Oja 1989; Sanger 1989; Miller 1988; Hancock, Baddeley, and Smith 1992; Földiák 1990). Often this has been accompanied by information-theoretic arguments. The first work along these lines was by Linsker (1988), who first proposed the "infomax" principle that underlies our own work. Linsker's approach, and that of Atick and Redlich (1990), Bialek, Ruderman, and Zee (1991), and van Hateren (1992) uses the second-order (covariance matrix) approximation of the required information-theoretic quantities, and generally assumes Gaussian signal and Gaussian noise, in which case the second-order information is complete. The explicit noise model and the restriction to second-order statistics mark the two differences between these approaches and our approach to infomax.

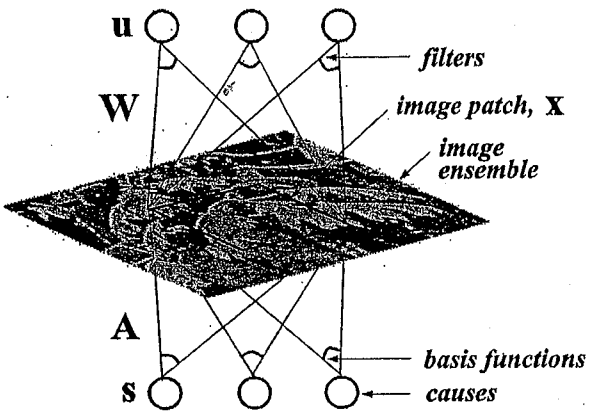

Figure 19.6
The blind linear image synthesis model (Olshausen and Field 1997). Here each patch, **x**, of an image is viewed as a linear combination of several (here three) underlying basis functions, given by the matrix **A**, each associated with an element of an underlying vector of "causes," **s**. Causes are viewed here as statistically independent "image sources." The causes are recovered (in a vector **u**) by a matrix of filters, **W**, which attempt to invert the unknown mixing of unknown basis functions constituting image formation.

The assumption of a noise model has been generally thought to be a necessary ingredient. In the case where the decorrelating filters are of the local ZCA type, the noise model is required (Atick and Redlich 1990) to avoid center-surround receptive fields with peaks a single pixel wide, as in figure 19.3, panel b (see also Atick and Redlich 1993). In the case of the PCA-style global filters, noise is automatically associated with the filters with high spatial frequency selectivity whose eigenvectors have small eigenvalues.

In both cases, it is questionable whether such assumptions about noise are useful. In the case of PCA, there is no a priori reason to associate signal with low spatial frequency and noise with high spatial frequency or, indeed, to associate signal with high amplitude components and noise with low amplitude. On the contrary, sharp edges, presumably of high interest, contain many high-frequency, low-amplitude components. In the case of local ZCA-type filters, some form of spatial integration is assumed necessary to average out photon shot noise. Yet we know photoreceptors and the brains associated with them can operate in the single-photon detection regime. Therefore shot noise is, in at least some cases, not considered by neural systems to be something noisy to be ignored, and such systems appear to operate at the limit of the spatial acuity allowed by their lattices of receptors.

In a general information-theoretic framework, there is nothing to distinguish signal and noise a priori, and we therefore question the use of the concept of noise in these models. Of course there are signals of lesser or greater relevance to an organism, but there is no signature in their spatial or temporal structure that distinguishes them as important or not. It is more likely that signal and noise are subjective concepts having to do with the prior expectations of the organism (or neural subsystem). In the case of the simple linear mappings we are considering, there is no internal state (other than the filters themselves) to store such prior expectations, and therefore we consider "noiseless infomax" to be the appropriate framework for making the first level of predictions based on information-theoretic reasoning.

The second difference in earlier infomax models, the restriction to second-order statistics, has been questioned by Field (1987, 1994) and Olshausen and Field (1997). This has coincided with a general rise in awareness that simple Hebbian-style algorithms without special constraints are unable to produce local oriented receptive fields like those found in area V1 of visual cortex, but rather produce solutions of the PCA or ZCA type, depending on the constraint put on the decorrelating filter matrix, **W**.

The technical reason for this failure is that second-order statistics correspond to the amplitude spectrum of a signal (because the Fourier transform of the autocorrelation function of an image is its power spectrum, the square of the amplitude spectrum.) The remaining information, higher-order statistics, corresponds to the phase spectrum. The phase spectrum is what we consider to be the informative part of a signal, since if we remove phase information from an image, it looks like noise, while if we remove amplitude information (for example, with zero-phase whitening, using a ZCA transform), the image is still recognizable. Edges and what we consider "features" in images are "suspicious coincidences" in the phase spectrum: Fourier analysis of an edge consists of many sine waves of different frequencies, all aligned in phase where the edge occurred.

As in our conclusions about "noise," we feel that a more general information-theoretic approach is required, an approach taking account of statistics of all orders. Such an approach is sensitive to the phase spectra of the images, and thus to their characteristic local structure. These conclusions are borne out by the results of ICA, which demonstrate the emergence of local oriented receptive fields, which second-order statistics alone fail to predict.

Several other approaches have arisen to deal with the unsatisfactory results of simple Hebbian and anti-Hebbian schemes. Field (1987, 1994) emphasized, using some of Barlow's arguments (1989), that the goal of an image transformation should be to convert "higher-order redundancy" into "first order-redundancy." These arguments led Olshausen and Field (1997) to attempt to learn receptive fields by maximizing sparseness. In terms of our figure 19.6, they attempted to find receptive fields (which they identified with basis functions—the columns of our **A** matrix) that have underlying causes, **u** (or **s**), and are as sparsely distributed as possible. The sparseness constraint is imposed by a nonlinear function that pushes the activity of the components of **u** toward zero.

Thus the similarity of the results produced by Olshausen and Field's network and ours may be explained by the fact that both produce what are perhaps the sparsest possible u_i distributions, though by different means. In emphasizing sparseness directly, rather than an information theoretic criterion, Olshausen and Field do not force their "causes" to have low mutual information, or even to be decorrelated. Thus their basis function matrices, unlike ours, are singular, and noninvertible, making it difficult for them to say what the filters are that correspond to their basis functions. Recently, Lewicki and Olshausen (1999), working with overcomplete representations, have overcome these problems.

Our approach, on the other hand, emphasizes independence over sparseness. Examining figures 19.4 and 19.5, we see that our filter outputs are also very sparse. This is because infomax with a sigmoid nonlinearity can be viewed as an ICA algorithm with an assumption that the independent components have super-Gaussian PDFs. It is worth mentioning that an ICA algorithm without this assumption will find a few sub-Gaussian (low-kurtosis) independent components, though most will be super-Gaussian (Lee, Girolami, and Sejnowski 1999).

Sparseness, as captured by the kurtosis, is one projection index often mentioned in projection pursuit methods (Huber 1985), which look in multivariate data for directions with "interesting" distributions. Intrator (1992; see chapter 18), who pioneered the application of projection pursuit reasoning to feature extraction problems, used an index emphasizing *multimodal* projections, and connected it with the BCM (Bienenstock, Cooper, and Munro 1982) learning rule. Following up, Law and Cooper

(1994) and Shouval (1995) used the BCM rule to self-organize oriented and somewhat localized receptive fields on an ensemble of natural images.

The BCM rule is a nonlinear Hebbian/anti-Hebbian mechanism. The nonlinearity undoubtedly contributes higher-order statistical information, but it is less clear than in Olshausen's network or our own how the nonlinearity contributes to the solution.

Another principle, predictability minimization, has also been brought to bear on the problem by Schmidhuber, Eldracher, and Foltin (1996). This approach attempts to ensure independence of one output from the others by moving its receptive field away from what is predictable (using a nonlinear "lateral" network) from the outputs of the others. Finally, Harpur and Prager (1996) have formalized an inhibitory feedback network that also learns nonorthogonal oriented receptive fields.

19.4 Biological Significance

The simplest properties of classical V1 simple cell receptive fields (Hubel and Wiesel 1968) are that they are *local* and *oriented*. These are properties of the filters in figure 19.4, while failing to emerge (without external constraints) in many previous self-organizing network models (Linsker 1988; Miller 1988; Atick and Redlich 1993; Troyer et al. 1999). However, the transformation from retina to V1, from analog photoreceptor signals to spike-coding pyramidal cells, is clearly much more complex than the $\mathbf{W}_I$ matrix, with which we have been working.

Nonetheless, evidence supports a feedforward origin for the oriented properties of simple cells in the cat (Ferster et al. 1996). Also the ZCA filters approximate the static response properties of ganglion cells in the retina and relay cells in the lateral geniculate nucleus, which, to a first approximation, prewhiten inputs reaching the cortex.

If we were to accept $\mathbf{W}_I$ as a primitive model of the retinocortical transformation, then several objections might arise. One might object to the representation learned by the algorithm: the filters in figure 19.3 are predominantly of high spatial frequency, even though spatial frequencies have been found to spread over several octaves in cortex (Hubel and Wiesel 1974). The reason there are so many high spatial frequency filters is because they are smaller, therefore more are required to "tile" the 12×12 pixel array of the filter. However, active control of fovea-based eye movements and the topographic nature of V1 spatial maps means that visual cortex samples images in a very different way from our random, spatially unordered sampling of 12×12 pixel patches. Changing our model to make it more realistic in these two respects could produce different results.

Another important issue with regard to redundancy reduction is the significant redundancy across the encodings of neighboring image patches. The spatial decorrelation of natural images in a wavelet representation leads to suppressive interactions between filters in neighboring patches (Schwartz and Simoncelli 1999), similar to what has been reported in the primary visual cortex (Das and Gilbert 1999).

The approach taken here can also be extended to redundancy that occurs in sequences of images (van Hateren and Ruderman 1998). Here the inputs are three-dimensional spatiotemporal patterns and the filters have the properties of directionally selective simple cells found in the primary visual cortex.

The properties of neurons in the visual cortex depend on experience as well as genetically determined mechanisms, so it is natural to ask whether there are biological ways that an ICA algorithm could be implemented. Although the learning rule we used is nonlocal, it involves a feedback of information from, or within, the output layer. There are many

ways that such a biophysical self-organizational processes could be accomplished using local spatial media where the feedforward and the feedback of information are tightly functionally coupled (Bell 1992; Eagleman et al. 2001).

Regardless of whether any biological system implements an unsupervised learning rule such as ICA, the results allow us to interpret the response properties of simple cells in visual cortex as a form of redundancy reduction, as Barlow conjectured. Care must be taken, however, in drawing strong conclusions about visual cortical encodings, from models consisting of only a single static linear transformation.

19.5 Conclusion

What coding principles predict the formation of localized, oriented receptive fields? Barlow's answer was that edges are suspicious coincidences in an image. Based on the principles of information theory (Cover and Thomas 1991), Barlow proposed that our visual cortical feature detectors might be the end result of a redundancy reduction process (Barlow 1989; Atick 1992), in which the activation of each feature detector is as *statistically independent* from the others as possible.

We approached this problem through unsupervised learning in a single layer of linear filters based on an ensemble of natural images. The localized edge detectors that were produced have phase sensitivity as a result of the sensitivity of ICA to higher-order statistics.

Edges (or rather, areas of local contrast) are the first level of structure in images, being detectable by linear filters alone. The analogous cells in area V1, called "simple cells," are the last in the visual system to fit a "cardinal cell" model (von der Malsburg 1999)—that is, there is one cell for each location and type of object (i.e., orientation). Complex cells in area V1, which are somewhat location invariant, and neurons further up the visual processing pathways, which have many invariant properties, present a huge challenge to unsupervised learning models. Can their properties be predicted (or retrodicted) and their coding properties thus explained?

We believe the answer to this question is yes, and that it will involve the formulation of algorithms related to ICA, in which group-theoretic symmetries in probability distributions are identified with the subspaces in which they are embedded. Von der Malsburg has argued convincingly for many years that invariant coding and "feature binding" are the same problem, so we expect such learning algorithms will help bridge, in an information-theoretic way, the difficult gap between sensory and perceptual learning.

This will also greatly increase the computational power of abstract unsupervised learning techniques.