The 'Independent Components' of Natural Scenes are Edge Filters.

Anthony J. Bell and Terrence J. Sejnowski

Howard Hughes Medical Institute
Computational Neurobiology Laboratory
The Salk Institute
10010 N. Torrey Pines Road
La Jolla, California 92037*

Abstract

Field (1994) has suggested that neurons with line and edge selectivities found in primary visual cortex of cats and monkeys form a sparse, distributed representaton of natural scenes, and Barlow (1989) has reasoned that such responses should emerge from an unsupervised learning algorithm that attempts to find a factorial code of independent visual features. We show here that a new unsupervised learning algorithm that is based on information maximisation, a non-linear 'infomax' network (Bell and Sejnowski, 1995) when applied to an ensemble of natural scenes, produces sets of visual filters that are localised and oriented. Some of these filters are Gabor-like and resemble those produced by the sparseness-maximisation network of Olshausen & Field (1996). In addition, the outputs of these filters are as independent as possible, since the infomax network is able to perform Independent Components Analysis (ICA). We compare the resulting ICA filters and their associated basis functions, with other decorrelating filters produced by Principal Components Analysis (PCA) and zero-phase whitening filters (ZCA). The ICA filters have more sparsely distributed (kurtotic) outputs on natural scenes. They also resemble the receptive fields of simple cells in visual cortex, which suggests that these neurons form a natural, information-theoretic co-ordinate system for natural images.

*Please send comments to tony@salk.edu. This paper is submitted to *Vision Research.*

1 Introduction.

Both the classic experiments of Hubel & Wiesel (1968) on neurons in visual cortex, and several decades of theorising about feature detection in vision (Marr & Hildreth 1980), have left open the question most succinctly phrased by Barlow & Tolhurst (1992) "Why do we have edge detectors?". That is: are there any coding principles which would predict the formation of localised, oriented receptive fields? Barlow's answer is that edges are suspicious coincidences in an image. The mathematical framework for analysing such 'coincidences' is Information Theory (Cover & Thomas 1991). Barlow was thus led to propose 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 supposed to be as *statistically independent* from the others as possible. Such a 'factorial code' potentially involves dependencies of all orders, but most studies have used only the second-order statistics required for *decorrelating* the outputs of a set of feature detectors.

A variety of Hebbian feature-learning algorithms for decorrelation have been proposed (Linsker 1992, Miller 1988, Oja 1989, Sanger 1989, Foldiak 1990, Atick & Redlich 1993), but in the absense of particular external constraints the solutions to the decorrelation problem are non-unique (see Section 3). One popular decorrelating solution is Principal Components Analysis (PCA) but the principal components of natural scenes amount to a global spatial frequency analysis (Hancock et al 1992). Therefore, second order statistics alone do not suffice to predict the formation of localised 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 (Intrator 1992) with a 'Projection Pursuit' (Huber 1985) flavour, the most successful of which has been Olshausen & Field's (1996) demonstration of the self-organisation of local, oriented receptive fields using a sparseness criterion.

Here we present results similar to those of Olshausen & Field, using a direct information-theoretic criterion which maximises the joint entropy of a non-linearly transformed output feature vector. We have previously demonstrated the ability of this non-linear information maximisation process (Bell & Sejnowski 1995) to find statistically independent components to solve the problem of separating mixed audio sources (Jutten & Herault 1991). This 'Independent Components Analysis' (ICA) problem (Comon 1994) is equivalent to Barlow's redundancy reduction problem, therefore if Barlow's reasoning is correct, we would expect the ICA solution to yield localised edge detectors.

That it does so is the primary result of this paper. The secondary result is that the outputs of the resulting filters are indeed, 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's network from an information-theoretic point of view.

We will come back to the issues of sparseness, noise and higher-order statistics in the Discussion (Section 7). First, we describe more concretely the filter-learning problem. An earlier account of the application of these techniques to natural sounds appears in Bell & Sejnowski (1996).

2 Blind separation of natural images.

The starting point is that of Olshausen & Field 1996, depicted in Figure 1. 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:

$$\mathbf{x} = \mathbf{A}\mathbf{s}. \tag{1}$$

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:

$$\mathbf{u} = \mathbf{W}\mathbf{x} \tag{2}$$

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 (non-zero entries only on the diagonal), such a system has converged when:

$$\mathbf{u} = \mathbf{W}\mathbf{A}\mathbf{s} = \mathbf{P}\mathbf{S}\mathbf{s}. \tag{3}$$

The scaling and permuting of the causes are arbitrary, unknowable factors, so we will 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 which 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'. A number of proposals are discussed in Section 7, however in the next two sections, we concentrate on algorithms producing causes which are decorrelated, and those attempting to produce causes that are statistically independent.

3 Decorrelation and Independence.

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

$$\langle \mathbf{u}\mathbf{u}^T \rangle = \text{diagonal matrix} \tag{4}$$

In general, there will be many $\mathbf{W}$'s which decorrelate. For example, in the case of Eq.(2) when $\langle \mathbf{u}\mathbf{u}^T \rangle = \mathbf{I}$, then:

$$\mathbf{W}^T\mathbf{W} = \langle \mathbf{x}\mathbf{x}^T \rangle^{-1} \tag{5}$$

which clearly leaves freedom in the choice of $\mathbf{W}$. There are, however, several special solutions to Eq.(5).

1. The Orthogonal (global) Solution [$\mathbf{W}\mathbf{W}^T = \mathbf{S}$]. Principal Components Analysis (PCA) is the orthogonal solution to Eq.(4). The principal components come from the eigenvectors of the covariance matrix, which are the columns of a matrix, $\mathbf{E}$, satisfying:

$$\mathbf{E}\mathbf{D}\mathbf{E}^{-1} = \langle \mathbf{x}\mathbf{x}^T \rangle \tag{6}$$

where $\mathbf{D}$ is the diagonal matrix of eigenvalues. Substituting Eq.(6) into Eq.(5) and solving for $\mathbf{W}$ gives the PCA solution, $\mathbf{W}_P$:

$$\mathbf{W}_P = \mathbf{D}^{-\frac{1}{2}}\mathbf{E}^T \tag{7}$$

This solution is unusual in that the filters (rows of $\mathbf{W}_P$) are orthogonal, so that $\mathbf{W}\mathbf{W}^T = \mathbf{D}^{-1}$, a scaling matrix. These filters thus have several special properties: (1) The PCA filters define orthogonal directions in the vector space of the image. (2) The PCA basis functions (columns of $\mathbf{A}_P$, or rows of $\mathbf{W}_P^{-T}$ – see Fig.1) are just scaled versions of the PCA filters (rows of $\mathbf{W}_P$). This latter property is true because $\mathbf{W}\mathbf{W}^T = \mathbf{D}^{-1}$ means that $\mathbf{W}^{-T} = \mathbf{D}\mathbf{W}$. (3) 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 Fig.3a.

2. The Symmetrical (local) Solution [$\mathbf{W}\mathbf{W}^T = \mathbf{W}^2$]. If we force $\mathbf{W}$ to be symmetrical, so that $\mathbf{W}^T = \mathbf{W}$, then the solution, $\mathbf{W}_Z$ to Eq.(5) is:

$$\mathbf{W}_Z = \langle \mathbf{x}\mathbf{x}^T \rangle^{-1/2}. \tag{8}$$

Like most other decorrelating filters, and unlike PCA, the basis functions and the filters coming from $\mathbf{W}_Z$ will be different from each other, and neither will be orthogonal. We might call this solution ZCA since the filters it produces are zero-phase (symmetrical). ZCA is in several ways the polar opposite of PCA. It produces *local* (centre-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. $\mathbf{W}_Z$ is related to the transforms described by Goodall (1960) and Atick & Redlich (1993).

Example ZCA filters and basis functions are shown in Fig.3b.

3. The Independent (semi-local) Solution [$f_{\mathbf{u}}(\mathbf{u}) = \prod_i f_{u_i}(u_i)$]. Another way to constrain the solution is to attempt to produce outputs which are not just decorrelated, but statistically independent, the much stronger requirement of Independent Components Analysis, or ICA (Jutten & Herault 1991, Comon 1994). The u_i are independent when their probability distribution, $f_{\mathbf{u}}$, factorises as follows: $f_{\mathbf{u}}(\mathbf{u}) = \prod_i f_{u_i}(u_i)$, equivalently, when there is zero mutual information between them: $I(u_i, u_j) = 0$, $\forall i \neq j$. A number of approaches to ICA have some relations with the one we describe below, notably Cardoso & Laheld 1996, Karhunen et al 1996, Amari et al 1996, Cichocki et al 1994 and Pham et al 1992. We refer the reader to these papers, to the two above, and to Bell & Sejnowski (1995a) for further background on ICA.

As we will show, in Section 6, ICA on natural images produces decorrelating filters which are sensitive to both phase (locality) and frequency information, just as in transforms involving oriented Gabor functions (Daugman 1985) or wavelets[1]. They are thus semi-local, depicted in Fig.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 Fig.3d and their corresponding basis functions are shown in Fig.3e.

4 An ICA algorithm.

It is important to recognise 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, since 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.

The approach which we developed in Bell & Sejnowski (1995a) was to maximise by stochastic gradient ascent the joint entropy, $H[g(\mathbf{u})]$, of the linear transform squashed by a sigmoidal function, g. When the non-linear function is the same (up to scaling and shifting) as the cumulative density functions (c.d.f.s) of the underlying independent components, it can be shown (Nadal & Parga 1995)[2] that such a non-linear 'infomax' procedure also minimises the mutual information between the u_i, exactly what is required for ICA.

However, in most cases we must pick a non-linearity, g, without any detailed knowledge of the probability density functions (p.d.f.s) of the underlying independent components. The resulting 'mismatch' between the gradient of the nonlinearity used, and the underlying p.d.f.s may cause the infomax solution to deviate from an ICA solution. In cases where the p.d.f.s 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 tanh nonlinearities, that maximisation of $H[g(\mathbf{u})]$ still leads to ICA solutions, when they exist, as with our experiments on speech signal separation (Bell & Sejnoski 1995a). Although the infomax algorithm is described here as an ICA

[1] See the Proceedings of IEEE, 84, 4, April 1996 — a special issue on wavelets.

[2] In a previous conference paper (Bell & Sejnowski 1995b), we also published a proof of this result, which ought to have referenced the equivalent proof by Nadal & Parga.

algorithm, a fuller understanding needs to be developed of under exactly what conditions it may fail to converge to an ICA solution.

The basic infomax algorithm changes weights according to the entropy gradient. Defining $y_i = g(u_i)$ to be the sigmoidally transformed output variables, the learning rule is then:

$$\Delta \mathbf{W} \propto \frac{\partial H(\mathbf{y})}{\partial \mathbf{W}} = E\left[\frac{\partial \ln |J|}{\partial \mathbf{W}}\right] \tag{9}$$

In this, E denotes expected value, $\mathbf{y} = [g(u_1) \ldots g(u_N)]^T$, and $|J|$ is the absolute value of the determinant of the Jacobian matrix:

$$J = \det \left[\frac{\partial y_i}{\partial x_j}\right]_{ij} \tag{10}$$

In *stochastic* gradient ascent we remove the expected value operator in Eq.(9), and then evaluate the gradient to give (Bell & Sejnowski 1995):

$$\Delta \mathbf{W} \propto [\mathbf{W}^T]^{-1} + \hat{\mathbf{y}}\mathbf{x}^T \tag{11}$$

where $\hat{\mathbf{y}} = [\hat{y}_1 \ldots \hat{y}_N]^T$, the elements of which depend on the nonlinearity as follows:

$$\hat{y}_i = \frac{\partial}{\partial y_i}\frac{\partial y_i}{\partial u_i} = \frac{\partial}{\partial u_i} \ln \frac{\partial y_i}{\partial u_i} \tag{12}$$

Amari, Cichocki & Yang (1996) have proposed a modification of this rule which utilises the *natural* gradient rather than the *absolute* gradient of $H(\mathbf{y})$. The natural gradient exists for objective functions which are functions of matrices, as in this case, and is the same as the *relative* gradient concept developed by Cardoso & Laheld (1996). It amounts to multiplying the absolute gradient by $\mathbf{W}^T\mathbf{W}$, giving, in our case, the following altered version of Eq.(11):

$$\Delta \mathbf{W} \propto \frac{\partial H(\mathbf{y})}{\partial \mathbf{W}} \mathbf{W}^T\mathbf{W} = (\mathbf{I} + \hat{\mathbf{y}}\mathbf{u}^T)\mathbf{W} \tag{13}$$

This rule has the twin advantages over Eq.(11) of avoiding the matrix inverse, and of converging several orders of magnitude more quickly, for data, $\mathbf{x}$, that is not prewhitened. The speedup is explained by the fact that convergence is no longer dependent on the conditioning of the underlying basis function matrix,

$\mathbf{A}$, of Eq.(1). This is the equivariant property explained by Cardoso & Laheld (1996).

Writing Eq.(13) in terms of individual weights, we have:

$$\Delta w_{ij} \propto w_{ij} + \hat{y}_i \sum_k w_{kj} u_k \tag{14}$$

The weighted sum non-local term in this rule can be seen as the result of a simple backwards pass through the weights from the linear output vector, $\mathbf{u}$, to the inputs, $\mathbf{x}$, so that each weight 'knows the influence' of its input, x_j.

It is also possible to write the rule in recurrent terms. As in the well known Jutten-Herault (1991) network or that of Foldiak (1990), we may use a feedback matrix, $\mathbf{V}$, giving a network: $\mathbf{u} = \mathbf{x} - \mathbf{V}\mathbf{u}$. Solving this gives $\mathbf{u} = (\mathbf{I} + \mathbf{V})^{-1}\mathbf{x}$, showing that $\mathbf{V}$ is just a coordinate transform of the $\mathbf{W}$ of Eq.(2). The learning rule for $\mathbf{V}$ is therefore a coordinate transform of the rule for $\mathbf{W}$. This is calculated as follows. Since the relationship between $\mathbf{W}$ and $\mathbf{V}$ is $\mathbf{W} = (\mathbf{I} + \mathbf{V})^{-1}$, we may write $\mathbf{V} = \mathbf{W}^{-1} - \mathbf{I}$. Differentiating, and using the quotient rule for matrices gives:

$$\Delta\mathbf{V} = \Delta(\mathbf{W}^{-1}) = \mathbf{W}^{-1}(\Delta\mathbf{W})\mathbf{W}^{-1}. \tag{15}$$

Inserting Eq.(13) and rearranging gives the learning rule for a feedback weight matrix:

$$\Delta\mathbf{V} \propto (\mathbf{I} + \mathbf{V})(\mathbf{I} + \hat{\mathbf{y}}\mathbf{u}^T). \tag{16}$$

In terms of an individual feedback weight, v_{ij}, this rule is:

$$\Delta v_{ij} \propto \delta_{ij} + v_{ij} + u_j(\hat{y}_i + \sum_k v_{ik}\, \hat{y}_k) \tag{17}$$

where $\delta_{ij} = 1$ when $i = j$, otherwise 0. Thus the feedback rule is also non-local, this time involving a backwards pass through the (recurrent) weights, of quantities, $\hat{y}_k$, calculated from the *non-linear* output vector, $\mathbf{y}$. Such a recurrent ICA system has been further developed for recovering sources which have been linearly convolved with temporal filters by Torkkola (1996).

The non-locality of the algorithm is interesting when we come to consider the biological significance of the learned filters in Section 7.5.

5 Methods.

We took four natural scenes involving trees, leafs and so on[3] and converted them to greyscale byte values between 0 and 255. A training set, $\{\mathbf{x}\}$, was then generated of 17,595 12x12 samples from the images. The training set was 'sphered' by subtracting the mean and multiplying by two times the local symmetrical (zero-phase) whitening filter of Eq.(8):

$$\{\mathbf{x}\} \leftarrow 2\mathbf{W}_Z(\{\mathbf{x}\} - \langle \mathbf{x} \rangle). \tag{18}$$

This removes both first and second order statistics from the data, and makes the covariance matrix of $\mathbf{x}$ equal to $4\mathbf{I}$. This is an appropriately scaled starting point for further training since infomax (Eq.(13)) on raw data, with the logistic function, $y_i = (1 + e^{-u_i})^{-1}$, produces a $\mathbf{u}$-vector which approximately satisfies $\langle \mathbf{u}\mathbf{u}^T \rangle = 4\mathbf{I}$. Therefore, by prewhitening $\mathbf{x}$ in this way, we can ensure that the subsequent transformation, $\mathbf{u} = \mathbf{W}\mathbf{x}$, to be learnt should approximate an orthonormal matrix (rotation without scaling), roughly satisfying the relation $\mathbf{W}^T\mathbf{W} = \mathbf{I}$ (Karhunen et al 1996). This $\mathbf{W}$ moves the solution along the decorrelating manifold from ZCA to ICA (see Fig.2).

The matrix, $\mathbf{W}$, is then initialised to the identity matrix, and trained using the logistic function version of Eq.(13), in which Eq.(12) evaluates as: $\hat{y}_i = 1 - 2y_i$. The training was conducted as follows: 30 sweeps through the data were performed, at the end of each of which, the order of the data vectors was permuted to avoid cyclical behaviour in the learning. During each sweep, the weights were updated only every 50 presentations in order that the vectorised MATLAB code could be more efficient. The learning rate (proportionality constant in Eq.(13)) was set as follows: 21 sweeps at 0.001, and 3 sweeps at each of 0.0005, 0.0002 and 0.0001. This process took 2 hours running MATLAB on a Sparc-20 machine, though a reasonable result for 12x12 filters can be achieved in 30 minutes. To verify that the result was not affected by the starting condition of $\mathbf{W} = \mathbf{I}$, the training was repeated with several randomly initialised weight matrices, and also on data that was not prewhitened. The results were qualitatively similar, though convergence was much slower.

The full ICA transform from the raw image was calculated as the product of the sphering (ZCA) matrix and the learnt matrix: $\mathbf{W}_I = \mathbf{W}\mathbf{W}_Z$. The

[3] The images (gif files) used are available in the Web directory `ftp://ftp.cnl.salk.edu/pub/tony/VRimages`

basis function matrix, $\mathbf{A}$, was calculated as $\mathbf{W}_I^{-1}$. A PCA matrix, $\mathbf{W}_P$, was calculated from Eq.(7). The original (unsphered) data was then transformed by all three decorrelating transforms, and for each the kurtosis of each of the 144 filters was calculated, according to the formula:

$$K_i = \frac{\langle (u_i - \langle u_i \rangle)^4 \rangle}{\langle u_i^2 - \langle u_i^2 \rangle \rangle^2} - 3 \tag{19}$$

Then the mean kurtosis for each filter type (ICA, PCA, ZCA) was calculated, averaging over all filters and input data.

6 Results.

The filters and basis functions resulting from training on natural scenes are displayed in Fig.3 and Fig.4. Fig.3 displays example filters and basis functions of each type. The PCA filters, Fig.3a, 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, Fig.3c, or the original images, Fig.3d, are semi-local filters, most with a specific orientation preference. The basis functions, Fig.3e, calculated from the Fig.3d ICA filters, are not local, and naturally have the spatial frequency characteristics of the original images. Basis functions calculated from Fig.3d (as with PCA filters) are the same as the corresponding filters since the matrix $\mathbf{W}$ (as with $\mathbf{W}_P$) is orthogonal.

In order to show the full variety of ICA filters, Fig.4 shows, with lower resolution, all 144 filters in the matrix $\mathbf{W}$, in reverse order of the vector-lengths of the filters, so that the filters corresponding to higher-variance independent components appear at the top. The general result is that ICA filters are localised and mostly oriented. Unlike the basis functions displayed in Olshausen & Field (1996), they do not cover a broad range of spatial frequencies. However, the appropriate comparison to make is between the ICA basis functions, and the basis functions in Olshausen & Field's Figure 4. The ICA basis functions in Fig.3e are oriented, but not localised and therefore it is difficult to observe any multiscale properties. However, when we ran the ICA algorithm on Olshausen's images, which were preprocessed with a whitening/lowpass filter, our algorithm yielded basis functions which were localised multiscale Gabor patches qualitively similar to those in Olshausen's Figure 4. Part of the difference in our results is therefore attributable to different preprocessing techniques. Further discussion and comparison of these two approaches is deferred to Section 7.3.

Fig.5 shows the result of analysing the distributions (image histograms) produced by each of the three filter types. As empasized 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 Fig.5, where the log histograms are seen to be roughly linear across 12 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 Fig.6a,c,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, Fig.6b,d,f shows the corresponding distribution if the outputs of the filters were independent (ie: the outer product of the marginal (univariate) distributions in Fig.5). Only the ICA joint histogram captures well the 'diamond' $\diamond$-shape characteristic of the product of the sparse univariate distributions, thus satisfying to a greater extent, the independence criterion: $f_{u_1 u_2}(u_1, u_2) = f_{u_1}(u_1) f_{u_2}(u_2)$.

In summary, these simulations show that the filters found by the ICA algorithm of Eq.(13) with a logistic non-linearity are localised, oriented, and produce outputs distributions of very high kurtosis. The significance of these results is now addressed.

7 Discussion

A substantial literature exists on the self-organisation of visual receptive fields. Many contributions have emphasised the roles of decorrelation and PCA (Oja 1989, Sanger 1989, Miller 1988, Hancock et al 1992, Foldiak 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 which underlies our own work. Linsker's approach, and that of Atick & Redlich (1990), Bialek et al (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.

7.1 Noise.

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 (see Section 3), the noise model is required (Atick & Redlich 1990) to avoid centre-surround receptive fields with peaks a single pixel wide, as in Fig.3b (see also Atick & 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. On the contrary, sharp edges, presumably of high interest, contain many high frequency 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 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.

7.2 Second and higher-order statistics.

The second difference in earlier infomax models, the restriction to second-order statistics, has been questioned by Field (1987, 1994) and Olshausen & Field (1996). 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, $\mathbf{W}$.

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 recognisable. 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. This time, we mean 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 we report, which demonstrate the emergence of local oriented receptive fields, which second-order statistics alone fail to predict.

7.3 Sparseness.

Several other approaches have arisen to deal with the unsatisfactory results of simple Hebbian and anti-Hebbian schemes. Field (1987, 1994) emphasised, using some of Barlow's (1989) arguments, that the goal of an image transformation should be to convert 'higher-order redundancy' into 'first order-redundancy'. In formal terms, if the output of two filters is u_1 and u_2, we may write their joint entropy as the sum of their individual entropies, minus the mutual information between them:

$$H(u_1, u_2) = H(u_1) + H(u_2) - I(u_1, u_2) \tag{20}$$

What is meant by higher order redundancy here is the $I(u_1, u_2)$ term. The creation of 'Minimum Entropy codes' is the shifting of redundancy from the $I(u_1, u_2)$ term to the $H(u_1)$ and $H(u_2)$ terms. Assuming the $H(u_1, u_2)$ term to be constant, this minimisation of $I(u_1, u_2)$ creates minimum entropy in the marginal distributions. A low entropy for $H(u_1)$, for example, means that the distribution, $f_{u_1}(u_1)$, is *sparse*, and this quality is identified in Field (1994), with the fourth moment of the distribution, the kurtosis. Very sparse distributions are peaky with long tails, and have positive kurtosis. They are often referred to as 'super-gaussian'.

Field's arguments led Olshausen & Field (1996), in work that motivated our approach, to attempt to learn receptive fields by maximising sparseness. In terms of our Fig.1, they attempted to find receptive fields (which they identified with basis functions — the columns of our $\mathbf{A}$ matrix) which have underlying causes, $\mathbf{u}$ (or $\mathbf{s}$), which are as sparsely distributed as possible. The sparseness constraint is imposed by a non-linear function that pushes the activity of the components of $\mathbf{u}$ towards zero. This search for minimum entropy sparse codes does not guarantee the attainment of a factorial code (any more than our infomax net does), but the increase in redundancy of the u_i-distributions, while maintaining a full basis set, will, in general, remove mutual information from between the elements of $\mathbf{u}$.

Thus the similarity of the results produced by Olshausen'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 emphasising sparseness directly, rather than an information theoretic criterion, Olshausen & 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 non-invertible, making it difficult for them to say what the filters are that correspond to their basis functions. This is not a flaw, however. Presently, there is no reason why decorrelation or a full-rank filter matrix should be absolutely necessary properties of a neural coding system.

What is interesting to point out here, though, as shown in Fig.5 and Fig.6 is that the sought-after property of sparseness is an emergent property of a direct information theoretic approach.

Another exploration of a kurtosis-seeking network has been performed by Fyfe & Baddeley (1995), with slightly negative conclusions. In a further study, Baddeley (1996) argued against kurtosis-maximisation, partly on the grounds that it would produce filters which are two pixels wide. This is to some extent vindicated by our results in Fig.4, where the filters achieving the highest kurtosis in Fig.5 are seen to be dominated by very *thin* edge detectors. However whether such a result is 'unphysiological' is debatable (see section 7.5).

7.4 Projection pursuit and other approaches.

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) has pioneered the application of projection pursuit reasoning to feature extraction problems. He used an index emphasizing *multimodal* projections, and connected it with the BCM (Bienenstock, Cooper & Munro 1982) learning rule. Following up from this, Law & Cooper (1994) and Shouval (1995) used the BCM rule to self-organise oriented and somewhat localised receptive fields on an ensemble of natural images.

The BCM rule is a non-linear Hebbian/anti-Hebbian mechanism. The non-linearity 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 minimisation, has also been brought to bear on the problem by Schmidhuber (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 & Prager (1996) have formalised an inhibitory feedback network which also learns non-orthogonal oriented receptive fields.

7.5 Biological significance.

The simplest properties of classical V1 simple cell receptive fields (Hubel & Wiesel 1968), that they are *local* and *oriented*, are properties of the filters in Fig.4, while failing to emerge (without external constraints) in many previous self-organising network models (Linsker 1988, Miller 1988, Atick & Redlich

1993). However, the transformation from retina to V1, from analog photoreceptor signals to spike-coding pyramidal cells, is clearly much more complex than the matrix, $\mathbf{W}_I$, with which we have been working.

Nonetheless, recent evidence has been found for a feedforward origin to 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 arise. One might object to the representation learned by the algorithm: the filters in Fig.4 are predominantly of high spatial frequency, unlike the several-octave spread seen in cortex (Hubel & Wiesel 1974). The reason there are so many high spatial frequency filters is because they are smaller, therefore more are required to 'tile' the 12x12 pixel array of the filter. However, the active control of eye movements and the topographic nature of V1 spatial maps means that visual cortex samples images in a much different way from our random, spatially unordered sampling of 12x12 pixel patches. Changing our model to make it more realistic in these two respects could produce different results.

One might also judge the algorithm itself to be biologically implausible. The learning rule in Eq.(13) is non-local. The non-locality is less severe than the original algorithm of [10], which involved a matrix inverse. However, in both its feedforward Eq.(14) and feedback Eq.(17) versions, it involves a back-propagation of information from, or within, the output layer. One might try to imagine a mechanism capable of performing such a back-propagation. However, since it is difficult to identify the parameters of our static matrix, $\mathbf{W}_I$, with 'true' biophysical parameters, we prefer to imagine that potentially real biophysical self-organisational processes (see for example Bell (1992)) occur in local spatial media where the feedforward and the feedback of information are tightly functionally coupled, and where some microscopic and dynamic analogue of Eq.(13) may operate.

On a final note, what is notable about our learning rule is its deviation from the simple Hebbian/anti-Hebbian correlational way of thinking about unsupervised learning. There is a correlational component in Eq.(14), but it is between a nonlinearly transformed output, and a term which is a weighted feedback from the linear outputs. In the experimental search for biophysical learning mechanisms, perhaps too much focus has been given to simple correlational Hebbian rules.

7.6 Conclusion.

We have presented an analysis of the problem of learning a single layer of linear filters based on an ensemble of natural images. The localised edge detectors produced are the first such to result from an information theoretic learning rule, and their phase-sensitivity is a result of the sensitivity of our rule to higher-order statistics.

Edges are the first level of invariance in images, detectable by linear filters alone. Further levels of invariance (shifting, rotating, scaling, lighting) clearly exist with natural objects in natural settings. These further levels may be extractable using similar information theoretic techniques, but multi-layer non-linear networks will be necessary. If this can be done, it will greatly increase both the computational and the empirical predictive power of abstract unsupervised learning techniques.

Acknowledgements

This paper emerged through many extremely useful discussions with Bruno Olshausen and David Field. We are very grateful to them, and to Paul Viola and Barak Pearlmutter for other most helpful discussions. The work was supported by the Howard Hughes Medical Institute.

References

[1] Amari S. Cichocki A. & Yang H.H. 1996. A new learning algorithm for blind signal separation, *Advances in Neural Information Processing Systems 8*, MIT press.

[2] Atick J.J. & Redlich A.N. 1990. Towards a theory of early visual processing, *Neural Computation*, 2, 308-320

[3] Atick J.J. 1992. Could information theory provide an ecological theory of sensory processing? *Network* 3, 213-251

[4] Atick J.J. & Redlich A.N. 1993. Convergent algorithm for sensory receptive field development, *Neural Computation* 5, 45-60

[5] Baddeley R. 1996. Searching for filters with "interesting" output distributions: an uninteresting direction to explore?, *Network*, to appear

[6] Barlow H.B. 1989. Unsupervised learning, *Neural Computation* 1, 295-311

[7] Barlow H.B. & Tolhurst D.J. 1992. Why do you have edge detectors? *Optical society of America: Technical Digest*, 23, 172.

[8] Barlow H.B. 1994. What is the computational goal of the neocortex? in Koch C. (ed.) *Large-scale neuronal theories of the brain.* Cambridge Mass: MIT press.

[9] Bell A.J. 1992. Self-organisation in real neurons: Anti-Hebb in 'channel space?', in Moody J. et al (eds) *Advances in Neural Information Processing Systems 4*, 59-66, Morgan-Kaufmann 1992

[10] Bell A.J. & Sejnowski T.J. 1995a. An information maximization approach to blind separation and blind deconvolution, *Neural Computation*, 7, 1129-1159

[11] Bell A.J. & Sejnowski T.J. 1995b. Fast blind separation based on information theory, in *Proc. Intern. Symp. on Nonlinear Theory and Applications*, Las Vegas, Dec. 1995

[12] Bell A.J. & Sejnowski T.J. 1996. Learning the higher-order structure of a natural sound, *Network: Computation in Neural Systems*, 7, 2

[13] Bialek W. Ruderman D.L. & Zee A. 1991. Optimal sampling of natural images: a design principle for the visual system? in *Advances in Neural Information Processing Systems 1*, D. Touretzky (ed.), Morgan-Kaufmann

[14] Bienenstock E.L., Cooper L.N. & Munro P.W. 1982. Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex, *J. Neurosci.*, 2, 1, 32-48.

[15] Burr D.C. & Morrone M.C. 1990. Feature detection in biological and artificial vision systems, in Blakemore C. (ed.) *Vision: coding and efficiency*, Camb. Univ. Press

[16] Cardoso J-F. & Laheld B. 1996. Equivariant adaptive source separation, *IEEE Trans. on Signal Proc.*, to appear

[17] Cichocki A., Unbehauen R., & Rummert E. 1994. Robust learning algorithm for blind separation of signals, *Electronics Letters*, 30, 17, 1386-1387

[18] Comon P. 1994. Independent component analysis, a new concept? *Signal Processing*, 36, 287-314

[19] Cover T.M. & Thomas J.A. 1991. *Elements of information theory*, John Wiley.

[20] Daugman J.G. 1985. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters, *J. Opt. Soc. Am A*, 2, 7, 1160-1169

[21] Ferster D., Chung S. & Wheat H. 1996. Orientation selectivity of thalamic input to simple cells of cat visual cortex, *Nature*, 380, 249-252.

[22] 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, 2370-2393

[23] Field D.J. 1994. What is the goal of sensory coding? *Neural Computation* 6, 559-601

[24] Foldiak P. 1990. Forming sparse representations by local anti-Hebbian learning, *Biol. Cybern.*, 64, 165-170

[25] Fyfe C. & Baddeley R. 1995. Finding compact and sparse-distributed representations of visual images, *Network*, 6, 333-344

[26] Goodall M.C. 1960. Performance of stochastic net, *Nature*, 185, 557-558

[27] Hancock P.J.B., Baddeley R.J. & Smith L.S. 1992. The principal components of natural images, *Network*, 3, 61-72.

[28] Harpur G.F. & Prager R.W. 1996. Development of low entropy coding in a recurrent network, *Network*, to appear

[29] Haykin S. (ed.) 1994. *Blind Deconvolution*, Prentice-Hall, New Jersey.

[30] Hubel D.H. & Wiesel T.N. 1968. Receptive fields and functional architecture of monkey striate cortex, *J. Physiol.*, 195: 215-244

[31] Hubel D.H. & Wiesel T.N. 1974. Uniformity of monkey striate cortex: a parallel relationship between field size, scatter, and magnification factor. *J. Comp. Neurol.* 158, 295-306.

[32] Huber P.J. 1985. Projection pursuit, *The Annals of Statistics*, 13, 435-475

[33] Intrator N. 1992. Feature extraction using an unsupervised neural network, *Neural Computation*, 4, 98-107

[34] Jutten C. & Herault J. 1991. Blind separation of sources, part I: an adaptive algorithm based on neuromimetic architecture, *Signal processing* 24, 1-10

[35] Karhunen J., Wang L. & Jousensalo J. 1995. Neural estimation of basis vectors in Independent Component Analysis, *Proc. ICANN, Paris, 1995*

[36] Karhunen J., Oja E., Wang L., Vigario R. & Joutsenalo J. 1996. A class of neural networks for independent component analysis, submitted to *IEEE Trans. on Neural Networks.*

[37] Laughlin S. 1981. A simple coding procedure enhances a neuron's information capacity, *Z. Naturforsch.*, 36, 910-912

[38] Law C.C. & Cooper L.N. 1994. Formation of receptive fields in realistic visual environments according to the Bienestock, Cooper and Munro (BCM) theory, *Proc. Natl. Acad. Sci, USA*, 91, 7797-7801

[39] Linsker R. 1992. Local synaptic learning rules suffice to maximise mutual information in a linear network, *Neural Computation*, 4, 691-702

[40] Linsker R. 1988. Self-organization in a perceptual network. *Computer*, 21, 105-117

[41] Marr D. & Hildreth E. 1980. Theory of edge detection. *Proc. R. Soc. Lond. Ser. B*, 207. 187-217.

[42] Miller K.D. 1988. Correlation-based models of neural development, in *Neuroscience and Connectionist Theory*, M. Gluck & D. Rumelhart, eds., 267-353, Lawrence Erlbaum, Hillsdale, NJ

[43] Nadal J-P. & Parga N. 1994. Non-linear neurons in the low noise limit: a factorial code maximises information transfer. *Network*, 5, 565-581.

[44] Oja E. 1989. Neural networks, principal components and linear neural networks, *Neural Networks*, 5, 927-935.

[45] Olshausen B.A. & Field D.J. 1996. Natural image statistics and efficient coding, *Network: Computation in Neural Systems*, 7, 2.

[46] Pham D.T. Garrat P & Jutten C. 1992. Separation of a mixture of independent sources through a maximum likelihood approach, in *Proc. EUSIPCO*, p.771-774

[47] Ruderman D.L. 1994. The statistics of natural images, *Network: Computation in Neural Systems*, 5, 517-548

[48] Sanger T.D. 1989. Optimal unsupervised learning in a single-layer network, *Neural Networks*, 2, 459-473

[49] Schmidhuber J. Eldracher M. & Foltin B. 1996. Semi-linear predictability minimization produces well-known feature detectors, *Neural Computation*, to appear

[50] Shouval H. 1995. Formation and organisation of receptive fields, with an input environment composed of natural scenes, *Ph.D. thesis*, Dept. of Physics, Brown University.

[51] Torkkola K. 1996. Blind separation of convolved sources based on information maximisation, Proc. IEEE Workshop on Neural Networks and Signal Processing, Kyota, Japan, Sept. 1996

[52] van Hateren J.H. 1992. A theory of maximising sensory information, *Biol. Cybern.*, 68, 23-29

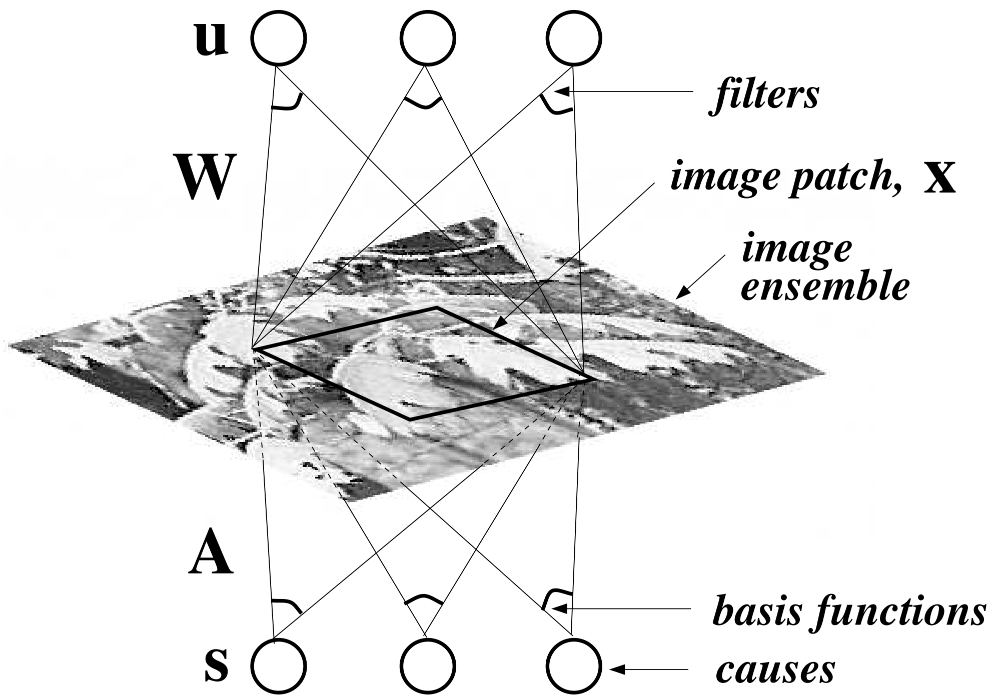

Figure 1: The Blind Linear Image Synthesis model (Olshausen & Field, 1996). Each patch, $\mathbf{x}$, of an image is viewed as a linear combination of several (here three) underlying basis functions, given by the matrix $\mathbf{A}$, each associated with an element of an underlying vector of 'causes', $\mathbf{s}$. In this paper, causes are viewed as statistically independent 'image sources'. The causes are recovered (in a vector $\mathbf{u}$) by a matrix of filters, $\mathbf{W}$, more loosely 'receptive fields', which attempt to invert the unknown mixing of unknown basis functions constituting image formation.

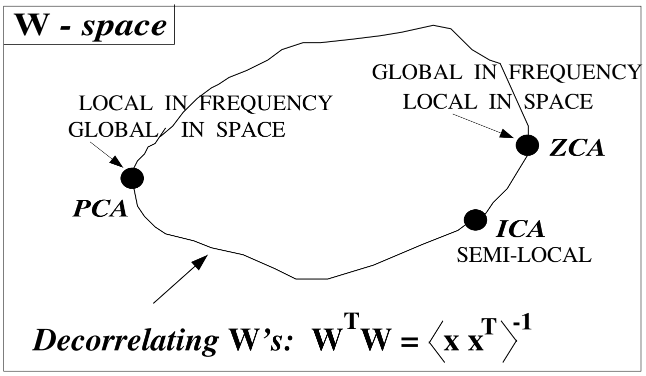

Figure 2: A schematic depiction of weight-space. A subspace of all matrices $\mathbf{W}$, here represented by the loop, has the property of decorrelating the input vectors, $\mathbf{x}$. On this manifold, several special linear transformations can be distinguished: PCA (global in space and local in frequency), ZCA (local in space and global in frequency), and ICA, a privileged decorrelating matrix which, if it exists, decorrelates higher as well as second order moments. ICA filters are localised, but not down to the single pixel level, as ZCA filters are (see Fig.3.)

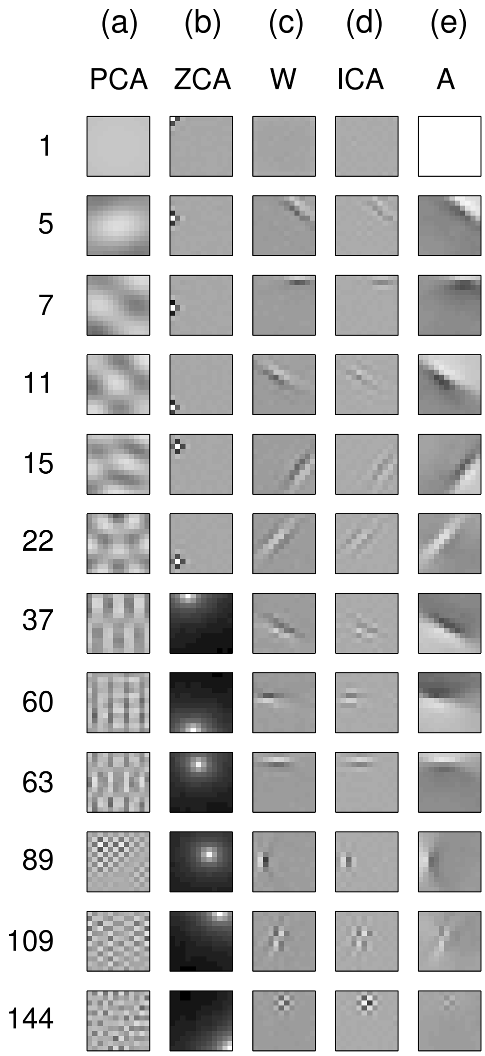

Figure 3: Selected decorrelating filters and their basis functions extracted from the natural scene data. Each type of decorrelating filter yielded 144 12x12 filters, of which we only display a subset here. Each column contains filters or basis functions of a particular type, and each of the rows has a number relating to which row of the filter or basis function matrix is displayed. (a) PCA ($\mathbf{W}_P$): The 1st, 5th, 7th etc Principal Components, calculated from Eq.(7), 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) ZCA ($\mathbf{W}_Z$): The first 6 entries in this column show the one-pixel wide centre-surround filter which whitens while preserving the phase spectrum. All are identical, but shifted. The lower 6 entries (37, 60 show the basis functions instead, which are the columns of the inverse of the $\mathbf{W}_Z$ matrix. (c) $\mathbf{W}$: the weights learnt by the ICA network trained on $\mathbf{W}_Z$-whitened data, showing (in descending order) the DC filter, localised oriented filters, and localised checkerboard filters. (d) $\mathbf{W}_I$: The corresponding ICA filters, calculated according to $\mathbf{W}_I = \mathbf{W}\mathbf{W}_Z$, looking like whitened versions of the $\mathbf{W}$-filters. (e) $\mathbf{A}$: the corresponding basis functions, or columns of $\mathbf{W}_I^{-1}$. These are the patterns which optimally stimulate their corresponding ICA filters, while not stimulating any other ICA filter, so that $\mathbf{W}_I\mathbf{A} = \mathbf{I}$.

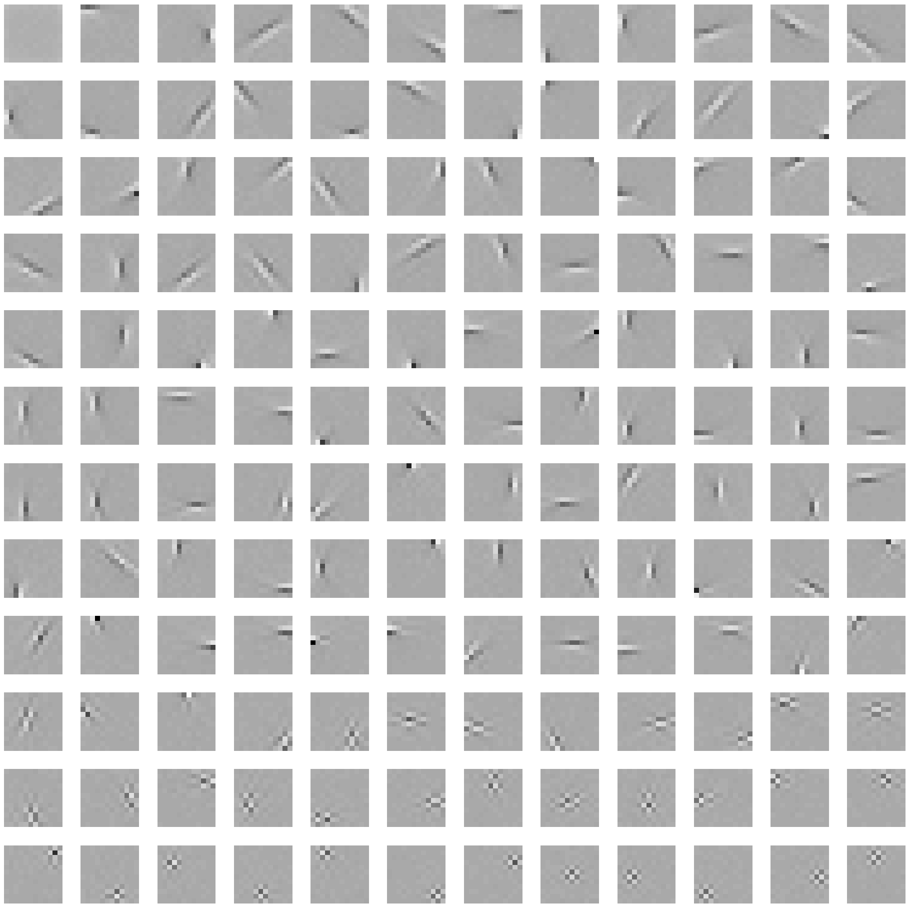

Figure 4: The matrix of 144 filters obtained by training on ZCA-whitened natural images. Each filter is a row of the matrix $\mathbf{W}$, and they are ordered left-to-right, top-to-bottom in reverse order of the length of the filter vectors. In a rough characterisation, and more-or-less in order of appearance, the filters consist of one DC filter (top left), 106 oriented filters (of which 35 were diagonal, 37 were vertical and 34 horizontal), and 37 localised checkerboard patterns. The diagonal filters are longer than the vertical and horizontal due to the bias induced by having square, rather than circular, receptive fields.

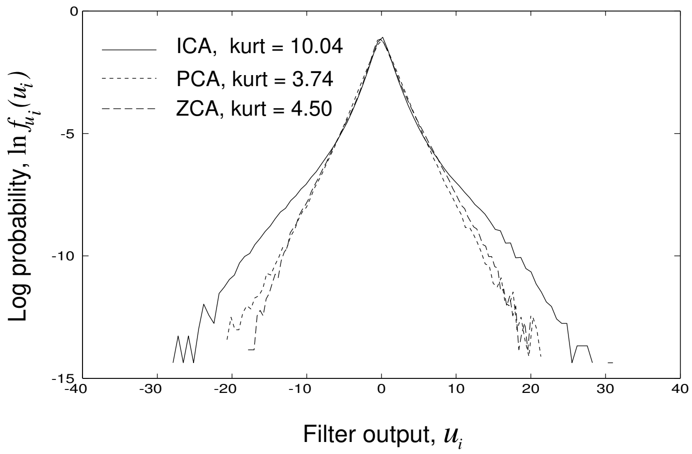

Figure 5: Log distributions of univariate statistics of the outputs of ICA, ZCA and PCA filters, averaged over all filters of each type. All three are approximately double-exponential distributions, but the more kurtotic ICA distribution, is slightly peakier and has a longer tail, showing that it is *sparser* than the others.

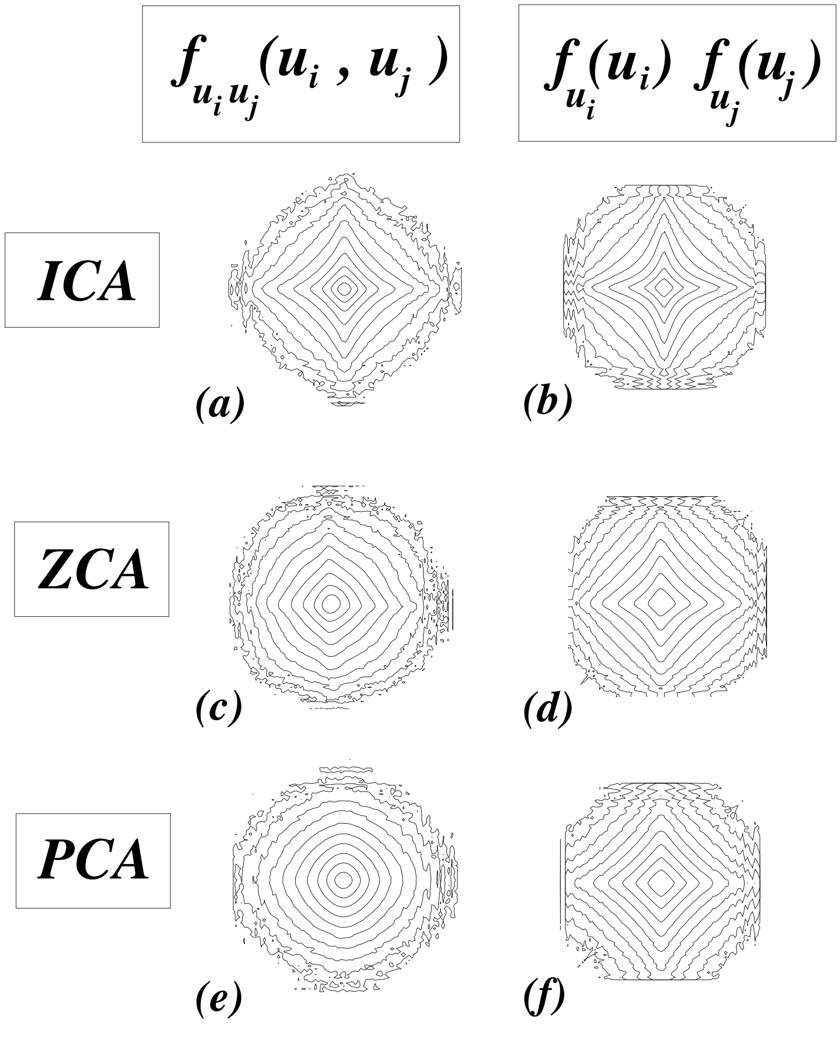

Figure 6: Contour plots of log distributions of pairwise statistics of the outputs of ICA, ZCA and PCA filters. Left column: joint log distributions averaged over all pairs of output filters of each type, and all images. Right column: product of marginal (univariate) distributions. The ICA solution best satisfies the independence criterion that the joint distribution has the same form as the product of the marginal distributions.