EXTENSIONS OF ICA AS MODELS OF NATURAL IMAGES AND VISUAL PROCESSING

Aapo Hyvärinen, Patrik O. Hoyer and Jarmo Hurri

Neural Networks Research Centre
Helsinki University of Technology
P.O. Box 5400, FIN-02015 HUT, Finland
http://www.cis.hut.fi/projects/compneuro/

ABSTRACT

Using statistical models one can estimate features from natural images, such as images that we see in everyday life. Such models can also be used in computional visual neuroscience by relating the estimated features to the response properties of neurons in the brain. A seminal model for natural images was linear sparse coding which, in fact, turned out to be equivalent to ICA. In these linear generative models, the columns of the mixing matrix give the basis vectors (features) that are adapted to the statistical structure of natural images. Estimated features resemble wavelets or Gabor functions, and provide a very good description of the properties of simple cells in the primary visual cortex. We have introduced extensions of ICA that are based on modelling dependencies of the "independent" components estimated by basic ICA. The dependencies of the components are used to define either a grouping or a topographic order between the components. With natural image data, these models lead to emergence of further properties of visual neurons: the topographic organization and complex cell receptive fields. We have also modelled the temporal structure of natural image sequences using models inspired by blind source separation methods. All these models can be combined in a unifying framework that we call bubble coding. Finally, we have developed a multivariate autoregressive model of the dependencies, which lead us to the concept of "double-blind" source separation.

1. INTRODUCTION

Recently, modeling image windows using statistical generative models has emerged as a new area of research, for reviews see [13, 17, 21]. Such an approach has applications both in image processing and visual neuroscience.

In image processing, using statistical generative models enables principled derivation of methods for denoising, compression, and other operations. In particular, a generative model gives a prior that can be used in Bayesian methods. In this paper, we will concentrate on applications in biological modelling, although the same models could be rather directly used in image processing.

A widely-spread assumption is that biological visual systems are adapted to process the particular kind of information they receive [3]. The visual system is important for survival and reproduction, and evolutionary forces thus drive the visual system towards signal processing that is optimal for the natural stimuli. This does not imply that genetic instructions completely determine the properties of the visual system: a large part of the adaptation to the natural stimuli could be accomplished during individual development.

Natural images have important statistical regularities that distinguish them from other kinds of input. For example, the gray-scale values or luminances at different pixels have robust and non-trivial statistical dependencies. Models of the statistical structure show what a statistically adapted representation of visual input should be like. Such models thus indicate what the visual system should be like if it followed the assumption of optimal adaptation to the visual input.

In the following, we first review very briefly the structure of the human visual system, see, e.g., [19] for a more detailed account. Then go on to discuss different models (based on ICA and related methods) that we and others have developed to model the statistics of natural images and the visual system.

2. HUMAN VISUAL SYSTEM

Figure 1 illustrates the earliest stages of the main visual pathway. Light is detected by the photoreceptors in the retinas, and the final output of the retinas is sent by the retinal ganglion cells through the optic nerve. The signal goes through the lateral geniculate nucleus (LGN)

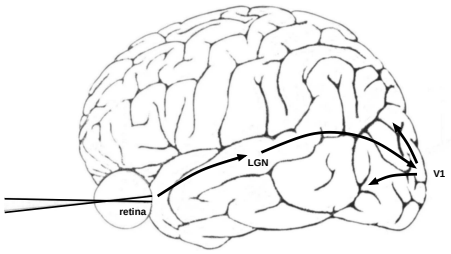

Figure 1: The main visual pathway in the human brain.

of the thalamus to the visual cortex at the back of the head, where most of the visual processing is performed.

The main information processing workload of the brain is carried by nerve cells, or neurons. The majority of neurons communicate by action potentials (also called spikes), which are electrical impulses traveling down the axons (something like wires) of neurons. Most research has concentrated on the neurons' *firing rates*, i.e. the number of spikes "fired" by a neuron per second (or some other time interval).

Thus, much of visual neuroscience has been concerned with measuring the firing rates of cells as a function of some properties of a visual input. For example, an experiment might run as follows: An image is suddenly projected onto a (previously blank) screen that an animal is watching, and the number of spikes fired by some recorded cell in the next second are counted. By systematically changing some properties of the stimulus and monitoring the elicited response, one can make a quantitative model of the response of the neuron. Such a model mathematically describes the response (firing rate) r_j of a neuron as a function of the stimulus $I(x, y)$.

In the early visual system, the response of a typical neuron depends only on the intensity pattern of a very small part of the visual field. This area, where light increments or decrements can elicit increased firing rates, is called the (classical) *receptive field* of the neuron. More generally, the concept also refers to the particular light pattern that yields the maximum response.

So, what light patterns actually elicit the strongest responses? This of course varies from neuron to neuron. The retinal ganglion cells as well as cells in the lateral geniculate nucleus typically have circular center-surround receptive field structure: Some neurons are excited by light in a small circular area of the visual field, but inhibited by light in a surrounding annulus. Other cells show the opposite effect, responding maximally to light that fills the surround but not the center. This is depicted in figure 2a.

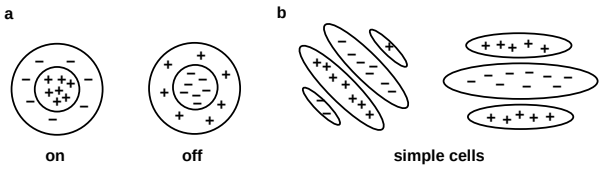

Figure 2: Typical classical receptive fields of neurons early in the visual pathway. Plus signs denote regions of the visual field where light causes excitation, minuses regions where light inhibits responses. (a) Retinal ganglion and LGN neurons typically exhibit center-surround receptive field organization, in one of two arrangements. (b) The majority of simple cells in V1, on the other hand, have oriented receptive fields.

The cells that we are modelling are in the primary visual cortex (V1). Cells in V1 have more interesting receptive fields. The so-called *simple cells* typically have adjacent elongated (instead of concentric circular) regions of excitation and inhibition. This means that these cells respond maximally to *oriented* image structure. This is illustrated in figure 2b.

Typically, classical receptive fields are modeled by a linear model: the response of a neuron can be reasonably predicted by a weighted sum of the image intensities, as in

$$r_j = \sum_{x,y} w_j(x, y) I(x, y), \tag{1}$$

where $w_j(x, y)$ contains the pattern of excitation and inhibition for light for the neuron j in question.

Although these linear models are useful in modeling many cells, there are also neurons in V1 called *complex cells* for which these models are completely inadequate. These cells do not show any clear spatial zones of excitation or inhibition. Complex cells respond, just like simple cells, selectively to bars and edges at a particular location and of a particular orientation; they are, however, relatively invariant to the spatial phase of the stimulus. An example of this is that reversing the contrast polarity (e.g. from white bar to black bar) of the stimulus does not markedly alter the response of a typical complex cell. The responses of complex cells have often been modeled by the classical 'energy model'. (The term 'energy' simply denotes the squaring operation.) In such a model we have

$$r_j = \left(\sum_{x,y} w_{j_1}(x, y) I(x, y) \right)^2 + \left(\sum_{x,y} w_{j_2}(x, y) I(x, y) \right)^2$$

where $w_{j_1}(x, y)$ and $w_{j_2}(x, y)$ are quadrature-phase Gabor functions, i.e., they have a phase-shift of 90 degrees, one being odd-symmetric and the other being

even-symmetric. It is often assumed that V1 complex cells pool the responses of simple cells, in which case the linear responses in the above equation are outputs of simple cells.

A further interesting point is how the receptive fields of neighboring cells are related. In the retina, the receptive fields of retinal ganglion cells are necessarily linked to the physical position of the cells. This is due to the fact that the visual field is mapped in an orderly fashion to the retina. Thus, neighboring retinal ganglion cells respond to neighboring areas of the visual field. However, there is nothing to guarantee the existence of a similar organization further up the visual pathway.

But the fact of the matter is that, just like in the retina, neighboring neurons in the LGN and in V1 tend to have receptive fields covering neighboring areas of the visual field. Yet this is only one of several types of organization. In V1, the orientation of receptive fields also tends to shift gradually along the surface of the cortex. In fact, neurons are often approximately organized according to several functional parameters simultaneously. This kind of *topographic organization* also exists in many other visual areas.

3. LINEAR MODELS OF NATURAL IMAGES

The statistical generative models in visual modelling are typically linear, or at least they incorporate a linear part. Let us denote by $I(x, y)$ the pixel gray-scale 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:

$$I(x, y) = \sum_{i=1}^{n} a_i(x, y)s_i \tag{2}$$

for all x and y. The s_i are stochastic coefficients, different from patch to patch.

In a neuroscientific interpretation, the latent variables s_i model the responses of simple cells, and the a_i are closely related to their receptive fields (see below). Thus, in the following, we will use the expressions "simple cell outputs" or "latent variables" interchangably.

For simplicity, we assume that the number of pixels equals the number of basis vectors, in which case the linear system in Eq. (2) can be inverted. Then, each latent variable or simple cell response is obtained by applying a linear transformation to the data; the linear transformation gives the receptive field. Denoting by w_i the coefficients of the transformation, the output of the simple cell with index i, when the input is an image patch I, is given by

$$s_i = \langle w_i, I \rangle = \sum_{x,y} w_i(x, y)I(x, y). \tag{3}$$

It can be shown [10] that the a_i are basically low-pass filtered versions of the receptive fields w_i. Therefore, the properties of the w_i and a_i are for most purposes identical.

Estimation of the model consists of determining the values of a_i, observing a sufficient number of patches I without observing the latent variables s_i. This is equivalent to determining the values of w_i, or the values of s_i for each image patch. The relation to ICA is now evident. If the latent variables s_i are assumed to be statistically independent and nongaussian, the linear generative model is nothing but the ICA model.

Just like in ICA, the estimation can be simplified by suitable preprocessing. First, we can consider only the local changes in gray-scale values (called contrast), and remove the local mean (called the DC component) from the image. This also implies that the s_i have zero mean. Second, we whiten the data in the spatial domain: The data is transformed to an image so that for any two spatial points (x, y) and (x', y') the value of $I(x, y)$ and $I(x', y')$ are uncorrelated, and all points are normalized to unit variance. In the whitened space, we can then consider orthonormal transformations only, i.e. $\sum a_i(x, y)a_j(x, y) = 0$ if $i \neq j$ and 1 if $i = j$.

Now, the question is: How to describe the statistical properties of natural images with the linear generative model? In other words, what are the statistical properties of linear transformations of the data? For example, are they nongaussian and independent enough to be modelled by ICA?

4. SPARSENESS AND ICA

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 supergaussianity or leptokurtosis [3, 13, 18], and to ICA estimation methods. The outputs of linear filters that model simple cell receptive fields are very sparse; in fact, they maximize a suitable defined measure of sparseness.

Sparseness is a property of a random variable. Sparseness means that the random variable takes very small (absolute) values and very large values more often than a gaussian random variable; to compensate, it takes values in between relatively more rarely. Thus, the random variable is activated, i.e. significantly nonzero, only rarely. This is illustrated in Fig. 3. We

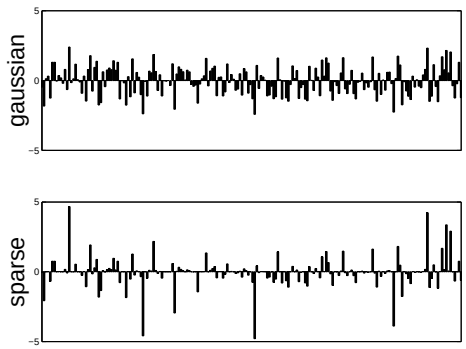

Figure 3: Illustration of sparseness. Random samples of a gaussian variable (top) and a sparse variable (bottom). The sparse variable is practically zero most of the time, occasionally taking very large values. Note that the variables have the same variance, and that the time structure is irrelevant in the definition of sparseness.

assume here and in what follows that the variable has zero mean.

The probability density function p of a sparse variable, say s, is characterized by a large value ("peak") at zero, and relatively large values far from zero ("heavy tails"). Here, "relatively" means compared to a gaussian distribution of the same variance. For example, the absolute value of a sparse random variable is often modelled as an exponential density. If the absolute value of a symmetric random variable has an exponential distribution, the distribution is called Laplacian. Scaling the distribution to have variance equal to one, the density function is then given by

$$p(s) = \frac{1}{\sqrt{2}} \exp(-\sqrt{2}|s|) \tag{4}$$

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, i.e. its second derivative is positive. Convexity implies that this expectation is large when s^2 typically takes values that are either very close to 0 or very large, i.e. when s is sparse.

For example, if G is the square function, sparseness is measured by the fourth moment $E\{s^4\}$. This is closely related to using the classical fourth-order cumulant called kurtosis, defined as $\mathrm{kurt}(s) = E\{s^4\} - 3(E\{s^2\})^2$. If the variance is normalized to 1, kurtosis is in fact the same as the fourth moment minus a constant (three). This constant is chosen so that kurtosis is zero for a gaussian random variable. If kurtosis is positive, the variable is called leptokurtic, which is a simple operational definition of sparseness.

However, kurtosis suffers from some adverse statistical properties [13], which is why in practice other functions may have to be used. Both information-theoretic and estimation-theoretic considerations show that in some ways the ideal functions would be such that $G(s^2)$ is equal to the logarithm of a sparse probability density function, optimally of s itself [13]. Then, the measure of sparseness is in fact essentially the same as entropy, or likelihood of the ICA model.

For example, taking the logarithm of the Laplacian density, one obtains

$$G(s^2) = -\alpha\sqrt{s^2} + \beta = -\sqrt{2}|s| - \log\sqrt{2} \tag{5}$$

In practice, a smoother version of the absolute value may be useful because the peak of absolute value at zero may lead to technical problems in optimization algorithms. A widely-used smoother version is given by $G(s^2) = \log\cosh\sqrt{s^2} = \log\cosh s$.

Maximization of sparseness with these sparseness measures is, in fact, very closely related to maximization of likelihood, or maximization of the negentropies of the estimated independent components. Assuming that the data is whitened, and the basis vectors are assumed orthonormal, the resulting method is nothing but basic ICA. The result of applying ICA on natural image patches [1, 13, 24] is shown in Figure 5.

5. TEMPORAL COHERENCE

An alternative to sparseness is given by temporal coherence or stability [4, 14, 6, 23, 25]. This means that when the input consists of natural image *sequences*, i.e. video data, the outputs of simple cells in subsequent time points should be "coherent" or "stable", i.e. change as little as possible. The change can be defined in many ways, and therefore temporal coherence can give rise to quite different definitions and measures.

First, it must be noted that using ordinary *linear* (auto)correlation or covariance is *not* enough to produce well-defined receptive fields. That is, if we measure the temporal coherence of a cell output $s(t)$, centered to have zero mean, as

$$\mathrm{corr}(s(t), s(t-\tau)) = E\{s(t)s(t-\tau)\}, \tag{6}$$

where τ is a time lag (delay), maximization of this measure does not characterize most simple cell receptive fields. In fact, this measure is maximized by low-pass filters, such as the DC component of image patches [6].

This failure of linear measures can be partly explained by basic results in the literature of blind source separation [13]. The autocovariance (for a given time lag) of the sum $a_i s_i + a_j s_j$ of two independent signals is given by $a_i^2 \mathrm{cov}(s_i(t), s_i(t-\tau)) + a_j^2 \mathrm{cov}(s_j(t), s_j(t-\tau))$. Consider a case where s_i and s_j have equal variances and autocovariances. Then, if the mixing coefficients fulfill $a_i^2 + a_j^2 = 1$, the mixture has the same variance *and* the same autocovariance as the original signals. There is an infinite number of such sums, and thus we cannot tell them apart if we just look at the autocovariance (and variance). This shows that maximization of autocorrelation does not properly define linear filters, and we have to use nonlinear autocorrelations.

Thus, we must use some kind of *nonlinear temporal correlations*. We have proposed [6] that temporal coherence could be measured by the correlation of squares (energies):

$$\mathrm{corr}(s(t), s(t-\tau)) = E\{s(t)^2 s(t-\tau)^2\} \quad (7)$$

It was found that the typical simple cell receptive fields maximize this criterion, just like sparseness. This measure was inspired by recent advances in the theory of blind source separation, where it has been shown that the correlation of squares is a valid measure for blind source separation [8]. In fact, this method can be seen as a variant of the class of blind separation methods using *nonstationary variance* [16, 20].

Thus, when properly defined and measured, temporal coherence does provide an alternative to sparseness, leading to the emergence of principal simple cell receptive field properties from natural images. The result of applying temporal coherence on natural image sequences is shown in Figure 6.

6. DEPENDENCIES BETWEEN COMPONENTS

6.1. Definition and models

The third statistical property considers the relationships between the different latent components (outputs of simple cells), which will be denoted by $s_i, i = 1, \ldots, n$. When using sparseness or temporal coherence, the outputs of simple cells s_i are usually assumed independent, i.e. the value of s_j cannot be used to predict s_i for $i \neq j$. To go beyond this basic framework, we need to model the statistical dependencies of the linear filters, assuming that their joint distribution is dictated by the natural image input [22, 9, 11].

Note that again, we must consider *nonlinear* correlations. Linear correlations are not interesting in this respect because they can easily be set to zero by standard whitening procedures. In fact, in ICA estimation, the components are often constrained to be uncorrelated [13].

When probing the dependence of s_i and s_j, a simple approach would be to consider the correlations of some nonlinear functions, just as in the case of temporal coherence. In image data, the principal form of dependency between two simple cell outputs seems to be captured by the correlation of their energies, or squares s_i^2. This means that

$$\mathrm{cov}(s_i^2, s_j^2) = E\{s_i^2 s_j^2\} - E\{s_i^2\} E\{s_j^2\} \neq 0. \quad (8)$$

Here, we assume that this covariance is positive, which is the usual case.

Intuitively, correlation of energies means that the cells tend to be active, i.e. 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. For example, if the variables are defined as products of two independent components o_i, o_j and a common "variance" variable v [11, 21]:

$$s_i = o_i v \quad (9)$$

$$s_j = o_j v \quad (10)$$

then s_i and s_j are uncorrelated, but their energies are not.

While the formulation above makes energy correlation easy to understand by using a separate variance variable v, it is not very suitable for practical computations, in which we need a simple expression for the joint probability density function of s_i and s_j. A simple density that incorporates both energy correlation and sparseness is given by [9, 11]

$$p(s_i, s_j) = \frac{2}{3\pi} \exp(-\sqrt{3}\sqrt{s_i^2 + s_j^2}) \quad (11)$$

This could be considered as a two-dimensional generalization of the Laplacian distribution, since it corresponds to a one-dimensional density where the exponential term would be proportional to $\exp(-\sqrt{3}\sqrt{s^2})$, which is as in Eq. (4) up to some scaling constants. (This density has been standardized to that its mean is zero and the variances are equal to one.) The correlation of energies in this probability distribution is illustrated in Fig. 4. A generalization of the probability density to more than two dimensions is straightforward by just taking the sum of the squares inside the square root in the exponential; the scaling and additive constants are then difficult to calculate but they are rarely needed.

Just as in the case of sparseness measures, the density in Eq. (11) gives us a measure of the combination of energy correlation and sparseness by considering the

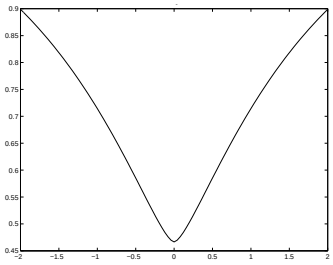

Figure 4: Illustration of the energy correlation in the probability density in Equation (11). The conditional variance of s_j (vertical axis) for a given s_i (horizontal axis). Here we see that conditional variance grows with the square (or absolute value) of s_i.

expectation of the log-density. We can take the logarithm of the density to obtain a function of the form

$$E\{G(s_i^2 + s_j^2)\} \tag{12}$$

where $G(b) = -\sqrt{b}$, up to irrelevant constants. This is a measure that is simple to compute. To get insight to this measure, consider what happens when G is the square function. Then the measure gives $E\{s_i^4 + s_j^4 + 2s_i^2 s_j^2\}$. The expectations of the first two terms measure sparseness just as kurtosis, while the expectation of the third is just the first term in the covariance of the squares. In practice, however, we prefer the logarithm of the density function to the square function because of the same statistical reasons (discussed above) that we prefer the log-density to kurtosis as a measure of sparseness.

6.2. Subspaces based on dependencies

The correlation of energies could be embedded in a model of natural image statistics in many ways. A very simple way would be to *divide the latent variables into groups* [2], 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 [9], it was found that the groups (called "independent subspaces") show emergence of complex cells properties, see Fig. 7. The sum of squares inside a group (which could be considered an estimate of the variance variable associated with that group) has the principal invariance properties of complex cells. Thus, simple cells that pool to the same complex cell have energy correlations, whereas simple cells that are not pooled together are independent.

6.3. Topography based on dependencies

Instead of a simple grouping, we have also proposed a more sophisticated way of modelling the correlations of squares of simple cell outputs, based on topography or spatial organization of the cells [11, 10]. The concept of cortical topography was reviewed earlier in section 2.

Let us assume that the components s_i are arranged on a two-dimensional grid or lattice as is typical in topographic models [15]. The restriction to 2D is motivated by cortical anatomy, but is not essential. The topography is formally expressed by a neighbourhood function $h(i, j)$ that gives the proximity of the components (cells) with indices i and j. (Note that these indices are two-dimensional). Typically, one defines that $h(i, j)$ is 1 if the cells are sufficiently close to each other, and 0 otherwise.

Our purpose was to 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. Since we are using the correlation of energies as the measure of dependency, the energies are strongly positively correlated for neighbouring cells.

We have defined such a statistical model, topographic ICA [11, 10], which incorporates just this kind of dependencies and can be estimated for natural images. When the model is applied on natural image data (see Fig. 8), the organization of simple cells is 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. This is the first model that shows emergence of all these principal properties of cortical topography [10].

An interesting point is that the topography defined by dependencies is closely related to complex cells: The topographic matrix $h(i, j)$ can be interpreted as the pooling weights from simple cell to complex cells. The pooling weights have now been set by making the assumption that complex cells only pool outputs of simple cells that are near-by on the topographic grid. Thus, we see how modelling the dependencies by topography is a generalization of a simple division of the cells into groups. Finally, note that a model of topography defined by energy correlation is very different from those typically used in models of (cortical) topography. Usually, the similarity of simple cells is defined by Euclidean distances or related measures, but correlation of energies is a strongly non-Euclidean measure.

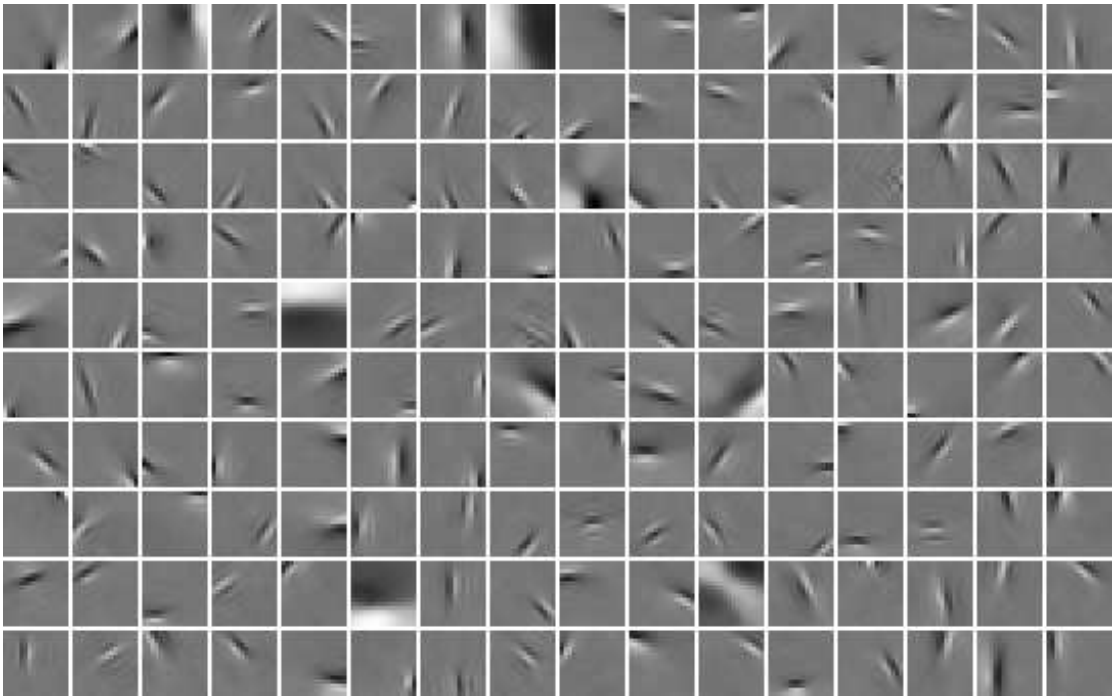

Figure 5: Basis vectors estimated by ICA or sparse coding.

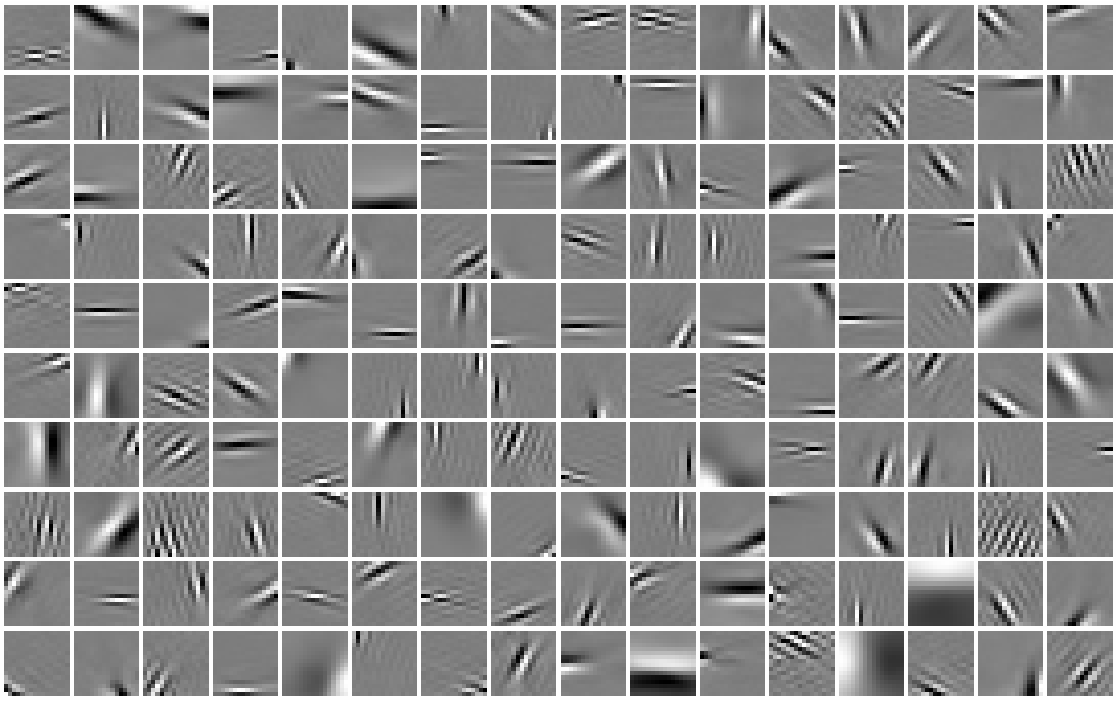

Figure 6: Basis vectors estimated by temporal coherence.

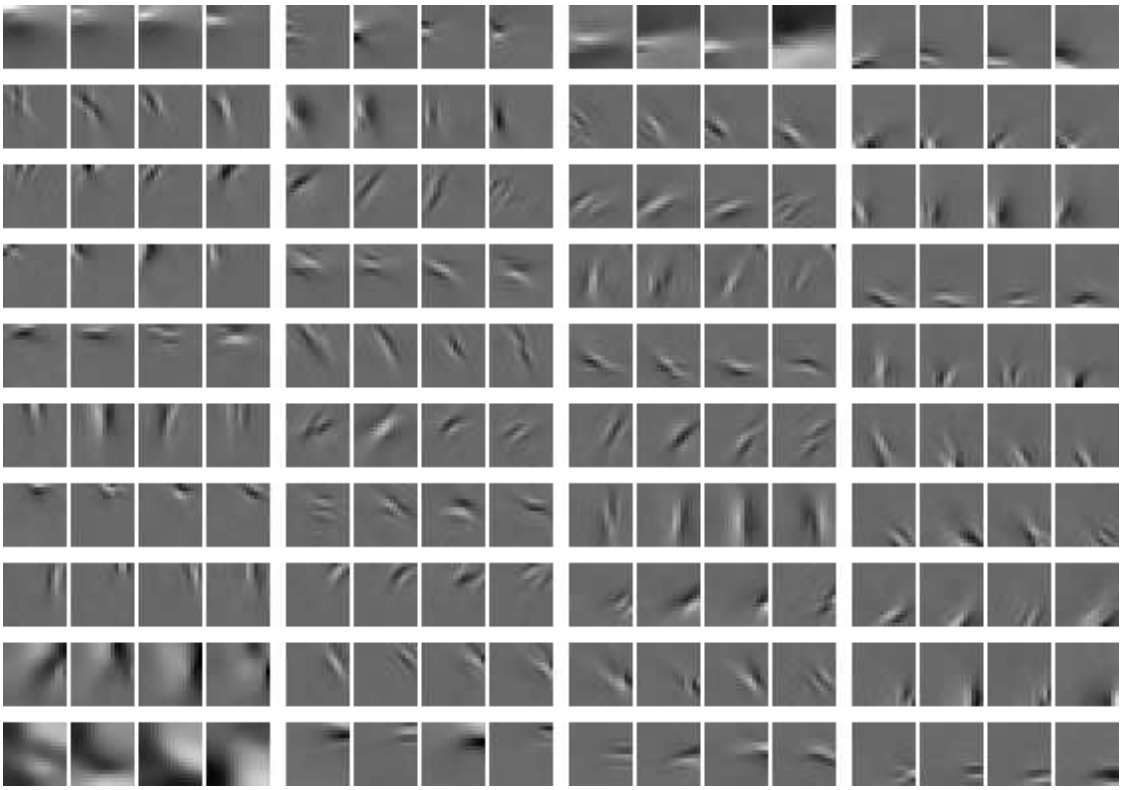

Figure 7: Basis vectors, and their grouping into 4-D subspaces, estimated by independent subspace analysis.

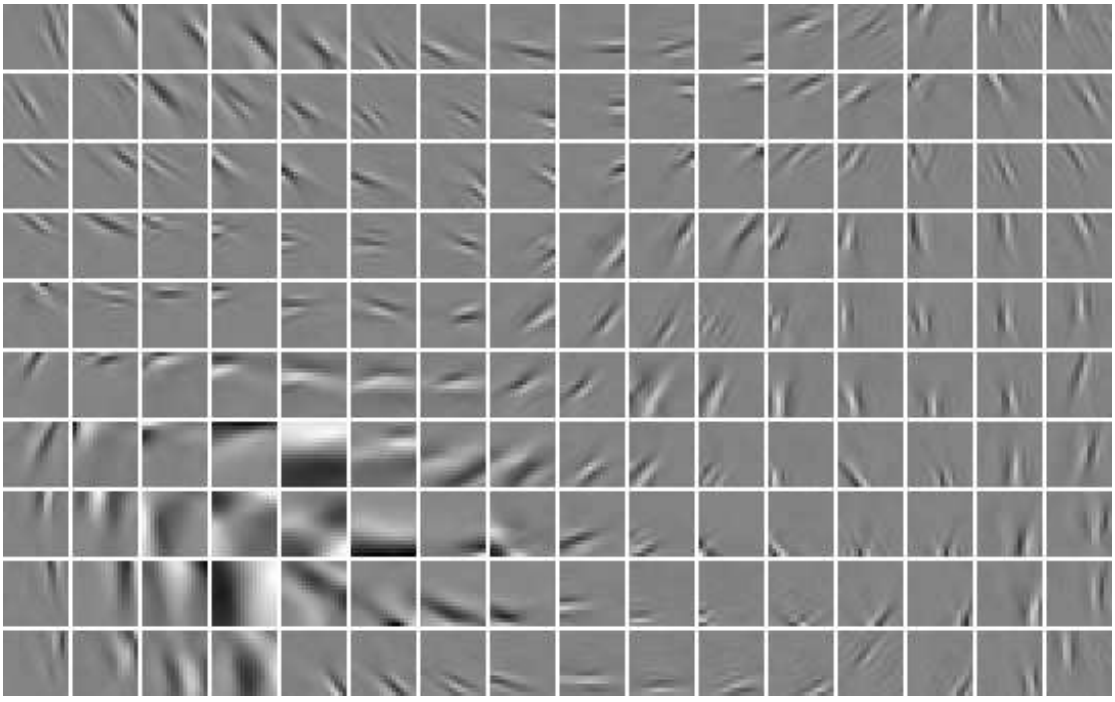

Figure 8: Basis vectors, and their topographic organization, estimated by topographic ICA.

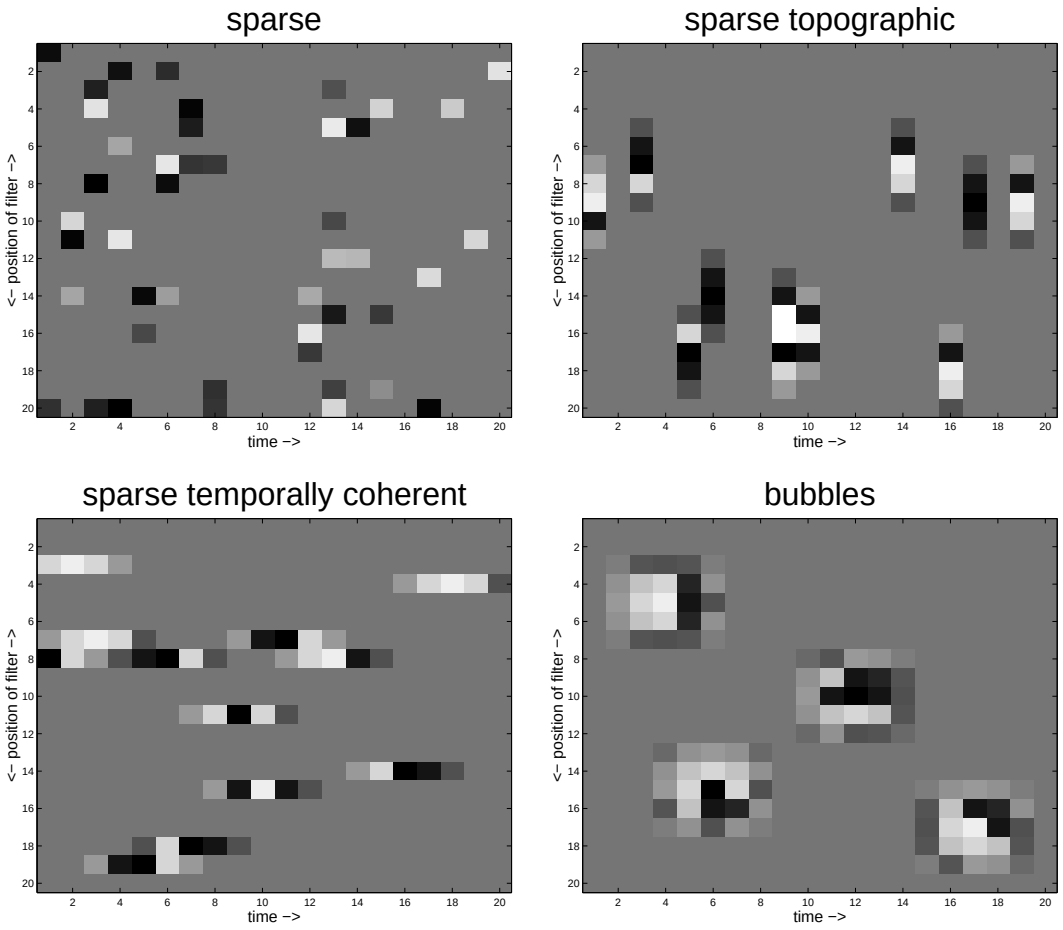

Figure 9: The four types of representation. The plots show the outputs of filters as a function of time and the position of the filter on the topographic grid. Each pixel is the activity on one unit at a give time point, gray being zero, white and black meaning positive and negative outputs. For simplicity, the topography is here one-dimensional. In the basic sparse (ICA) representation, the filters are independent. In the topographic representation, the activations of the filters are also spatially grouped. In the representation that has temporal coherence, they are temporally grouped. The bubble representation combines all these aspects, leading to spatiotemporal activity bubbles. Note that the two latter types of representation require that the data has a temporal structure, unlike basic sparse coding or ICA.

7. BUBBLES: A UNIFYING FRAMEWORK

Now we introduce 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 [12].

The key to the unifying framework is to note that in the models above, we used the same kind of dependence though variances (which expresses itself in the correlation of squares or energies) to model two different things: dependencies between the latent variables, and temporal dependencies of a single latent variable.

Combination of sparseness and topography means that each input activates a limited number of spatially limited "blobs" on the topographic grid, as in topographic ICA. If these regions are also temporally coherent, they resemble activity bubbles as found in many earlier neural network models. A spatiotemporal activity bubble thus means the *activation of a spatially and temporally limited region* of cells, or in general, representational units. This is illustrated in Fig. 9 for a one-dimensional topography.

What could such bubbles represent in practice? Since we are about to define a general-purpose unsupervised learning procedure, the meaning of bubbles depends on the data on which they are applied. In the case of natural image sequences, we can assume that the topographic grid is rather similar to the one obtained by topographic ICA. Then, a bubble would mean activation of Gabor-like basis vectors with similar orientation and spatial frequency, in near-by points on the image. This would correspond to a *short contour element* of given orientation and spatial frequency. In contrast to an "independent component" of an image, this contour element can move a bit, and its phase can change, during the temporal extent of the bubble.

Now, we formulate a generative model based on activity bubbles. We postulate a higher-order random process u that determines the variance at each point. This non-negative, highly sparse random process obtains independent values at each point in time and space (space referring to the topographic grid). For simplicity, let us denote the location on the topography by a single index i. Then, the variances v of the observed variables are obtained by a spatiotemporal convolution

$$v_i(t) = \sum_j h(i,j)[\varphi(t) * u_j(t)] \tag{13}$$

where $h(i,j)$ is the neighbourhood function that defines the spatial topography, and φ is a temporal smoothing kernel. The simple cell outputs are now obtained by multiplying simple gaussian white noise $o_i(t)$ by this variance signal:

$$s_i(t) = v_i(t)o_i(t) \tag{14}$$

Finally, the latent signals $s_i(t)$ are mixed linearly to give the image. Denote by $I(x,y,t)$ an image sequence. Then the mixing can be expressed as

$$I(x,y,t) = \sum_{i=1}^{n} a_i(x,y)s_i(t). \tag{15}$$

The three Eqs. (13–15) define a statistical generative model for natural image sequences.

The higher-order process $u_i(t)$ could be called the bubble process. When this process obtains a value that is different from zero, which is a rare event by definition, a bubble is created: The non-zero value spreads to neighbouring temporal and spatial locations due to the smoothing by φ and h. The spread of activation means that simple cells are have large variances inside that spatiotemporal window.

For experiments and estimation methods regarding the bubble model, see [12].

8. A TWO-LAYER MODEL WHERE BOTH LAYERS ARE ESTIMATED

We have also developed a two-layer model of natural image sequences that has the interesting property that both layers can be estimated [7]. This is in stark contrast to the models discussed above that fix the second layer (pooling of simple cell responses) beforehand, and only estimate the basis vectors (linear mixing matrix).

Technically, the estimation of two-layer models is quite difficult. In the models introduced above, estimation of the pooling weights is possible, in principle, by considering them as parameters just as the basis vectors. However, this introduces a normalization constant in the likelihood, because the integral of the probability density must equal one for any values of the pooling weights. Evaluation of this constant is most difficult.

We have been able to circumvent this problem by using a multivariate autoregressive model on the activity levels of simple cells. The activity levels correspond to the variances used in earlier sections, but for technical reasons, they are here defined simply as the absolute values. Let us denote by $\mathbf{abs}(\mathbf{s}(t))$ a vector that contains the absolute values of the elements of $\mathbf{s}(t)$. Further, let $\mathbf{v}(t)$ denote a driving noise signal in the autoregressive process. Let us denote by $\mathbf{M}$ a $K \times K$ matrix that gives the autoregressive coefficients, and let τ denote a time lag. Our model for the activities is a

constrained multidimensional first-order autoregressive process, defined by

$$\mathbf{abs}(\mathbf{s}(t)) = \mathbf{Mabs}(\mathbf{s}(t-\tau)) + \mathbf{v}(t). \quad (16)$$

Just as in ICA, the scale of the latent variables is not well defined, so we define that the variances of $s_i(t)$ are equal to unity.

The model is complicated by the fact that the absolute values must be non-negative. Thus, There are dependencies between the driving noise $\mathbf{v}(t)$ and the $\mathbf{s}(t-\tau)$. To define a generative model for the driving noise $\mathbf{v}(t)$ so that the non-negativity of the absolute values holds, we proceed as follows. Let $\mathbf{u}(t)$ denote a zero-mean random vector with components which are statistically independent of each other. We define $\mathbf{v}(t) = \max(-\mathbf{Mabs}(\mathbf{s}(t-\tau)), \mathbf{u}(t))$ where the maximum is computed component-wise.

To make the generative model complete, a mechanism for generating the signs of components $\mathbf{s}(t)$ must be included. We specify that the signs are generated randomly with equal probability for plus or minus after the strengths of the responses have been generated. All the signs are mutually independent, both over time and the cell population, and also independent of the activity levels. Note that one consequence of this random generation of signs is that that filter outputs are uncorrelated [7].

We have developed a method for estimating both the autoregressive matrix $\mathbf{M}$ and the basis vectors simultaneously. This is important because the set of basis vectors is not well-defined because of multiple local minima. Furthermore, there is little justification to assume that the maximally independent basis vectors given by ICA would be the optimal ones to use in a multi-layer model, since the structure of the higher layer affects the likelihood. For a description of the estimation method, and an interesting graphical representation of the resulting basis vectors and $\mathbf{M}$, see [7].

9. ESTIMATION THAT IS BLIND TO THE DEPENDENCIES

A most interesting result that we have obtained very recently is that the estimation method in the preceding section can be generalized to a model where the quantitative values of the dependencies (correlations of squares) are arbitrary. This leads to a separation method that is *double-blind* in the sense that no apriori assumptions are made either on the mixing matrix or on the higher-order correlations.

In the model we assume that the sources $s_i(t)$ have dependencies because the general activity levels, i.e. variances of the sources are not independent. Moreover, we assume that this activity levels change smoothly in time, as in methods based on nonstationary variance [16, 20, 8]. To model such dependencies, we assume, as above, that each source signal can be represented as a product of two random signals $v_i(t)$ and $o_i(t)$:

$$s_i(t) = v_i(t)o_i(t). \quad (17)$$

Here, $o_i(t)$ is an i.i.d. signal that is completely independent in time, and mutually independent over the index i as well. The dependencies (between the sources and over time) are only due the dependencies in $v_i(t)$, which is a non-negative signal giving the general activity level (variance). Thus, $v_i(t)$ and $v_j(t)$ are allowed to be statistically dependent. No particular assumptions on these dependencies are made, in order to have as blind a method as possible.

For simplicity, we use here the ordinary source separation notation and terminology. Assume that we observe an invertible linear transformation of the vectors of source signals:

$$\mathbf{x} = \mathbf{As} \quad (18)$$

with a square mixing matrix $\mathbf{A}$. The following theorem, whose proof will be found in a manuscript under preparation, shows how the model can be estimated without assumption on the form of the correlations of squares:

Theorem 1 *Assume the observed signals $x_i(t)$ are generated according to the model (18), and that the signals are preprocessed by spatial whitening to give the multidimensional signal $\mathbf{z}(t)$. Define the objective function:*

$$J(\mathbf{W}) = \sum_{i,j} \left[\mathrm{cov}(\mathbf{w}_i^T\mathbf{z}(t), \mathbf{w}_j^T\mathbf{z}(t-\tau))\right]^2 \quad (19)$$

where $\mathbf{W}$ is constrained to be orthogonal, and τ is a time lag. Assume that the matrix $\mathbf{K}$ defined by

$$\mathbf{K}_{ij} = \mathrm{cov}(s_i^2(t), s_j^2(t-\tau)) \quad (20)$$

is of full rank. Then, the objective function J is (globally) maximized when $\mathbf{WA}$ equals a signed permutation matrix, i.e. the $\mathbf{w}_i^T\mathbf{z}(t)$ equal the original sources $s_i(t)$ up to random signs.

10. CONCLUSION

Modelling the statistical structure of natural images is useful in vision research as well as in image processing. Possibly the most fundamental model is nothing but ICA, although it was originally motivated by sparse

coding. The obtained components are not really independent, which shows, in fact, an opportunity to model further aspects of the visual system.

We have developed models on the dependencies of the "independent" components. The most important kind of dependency seems to be the *correlation of squares* (energies), in other words, dependence through variances or activity levels. These dependencies are modelled by 1) independent subspaces and 2) a topographic organization of the components based on their dependency structure.

Further, we have modelled the temporal structure of natural image sequences using the very same kind of (temporal) dependencies through variances. This eventually lead to the unifying framework of spatiotemporal activity *bubbles*. Finally, we have developed a method of *double-blind* source separation, which is blind to the particular higher-order correlations of the components as well.

Future work will consist of extending this work to multi-layer models; see [5] for a first attempt.

11. REFERENCES

[1] A.J. Bell and T.J. Sejnowski. The 'independent components' of natural scenes are edge filters. *Vision Research*, 37:3327–3338, 1997.

[2] J.-F. Cardoso. Multidimensional independent component analysis. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'98)*, Seattle, WA, 1998.

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

[4] P. Földiák. Learning invariance from transformation sequences. *Neural Computation*, 3:194–200, 1991.

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

[6] J. Hurri and A. Hyvärinen. Simple-cell-like receptive fields maximize temporal coherence in natural video. *Neural Computation*. in press.

[7] J. Hurri and A. Hyvärinen. A two-layer temporal generative model of natural video exhibits complex-cell-like pooling of simple cell outputs. submitted.

[8] A. Hyvärinen. Blind source separation by nonstationarity of variance: A cumulant-based approach. *IEEE Transactions on Neural Networks*, 12(6):1471–1474, 2001.

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

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

[11] A. Hyvärinen, P. O. Hoyer, and M. Inki. Topographic independent component analysis. *Neural Computation*, 13(7):1527–1558, 2001.

[12] A. Hyvärinen, J. Hurri, and J. Väyrynen. Bubbles: A unifying framework for low-level statistical properties of natural image sequences. submitted.

[13] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley Interscience, 2001.

[14] C. Kayser, W. Einhäuser, O. Dümmer, P. König, and K. Körding. Extracting slow subspaces from natural videos leads to complex cells. In *Proc. Int. Conf. on Artificial Neural Networks (ICANN2001)*, pages 1075–1080, Vienna, Austria, 2001.

[15] T. Kohonen. *Self-Organizing Maps*. Springer, 1995.

[16] K. Matsuoka, M. Ohya, and M. Kawamoto. A neural net for blind separation of nonstationary signals. *Neural Networks*, 8(3):411–419, 1995.

[17] B. A. Olshausen. Principles of image representation in visual cortex. In L.M. Chalupa and J.S. Werner, editors, *The Visual Neurosciences*. MIT Press, 2003.

[18] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

[19] S. E. Palmer. *Vision Science – Photons to Phenomenology*. The MIT Press, 1999.

[20] D.-T. Pham and J.-F. Cardoso. Blind separation of instantaneous mixtures of non-stationary sources. In *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA2000)*, pages 187–193, Helsinki, Finland, 2000.

[21] E. P. Simoncelli and B.A. Olshausen. Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24:1193–216, 2001.

[22] E. P. Simoncelli and O. Schwartz. Modeling surround suppression in V1 neurons with a statistically-derived normalization model. In *Advances in Neural Information Processing Systems 11*, pages 153–159. MIT Press, 1999.

[23] J. Stone. Learning visual parameters using spatiotemporal smoothness constraints. *Neural Computation*, 8(7):1463–1492, 1996.

[24] J. H. van Hateren and A. van der Schaaf. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc. Royal Society, Ser. B*, 265:359–366, 1998.

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