Network: Computation in Neural Systems 7 (1996) 261–266. Printed in the UK

WORKSHOP PAPER

Learning the higher-order structure of a natural sound*

Anthony J Bell and Terrence J Sejnowski

Computational Neurobiology Laboratory, The Salk Institute, PO Box 85800, San Diego, CA 9186-5800, USA

Received 6 February 1996

Abstract. Unsupervised learning algorithms paying attention only to second-order statistics ignore the phase structure (higher-order statistics) of signals, which contains all the informative temporal and spatial coincidences which we think of as 'features'. Here we discuss how an Independent Component Analysis (ICA) algorithm may be used to elucidate the higher-order structure of natural signals, yielding their independent basis functions. This is illustrated with the ICA transform of the sound of a fingernail tapping musically on a tooth. The resulting independent basis functions look like the sounds themselves, having similar temporal envelopes and the same musical pitches. Thus they reflect both the phase and frequency information inherent in the data.

1. The poverty of second-order statistics

Natural signals have characteristic statistical dependencies across space and time. One view of sensory systems is that they must uncover these dependencies by processing them with filters whose form depends on the characteristic statistics of the ensemble of signals to which they are exposed (Barlow 1989, Atick and Redlich 1990). Considerable effort has gone into finding unsupervised learning algorithms able to self-organize to produce such filters (Linsker 1988, Miller 1988, Oja 1989, Sanger 1989, Foldiak 1990, Intrator 1992, Atick and Redlich 1993 and many others).

These efforts have been criticized by Field (1987, 1994). A major component of Field's argument is that the above methods are sensitive only to second-order statistics, since they all use correlation-based learning rules (i.e. Hebbian and/or anti-Hebbian rules). Most of the methods bear some relation to principal components analysis (the Karhunen–Loeve transform), a second-order statistical technique. The most informative features of natural signals, however, require higher-order statistics for their characterization. An edge in an image, or the transient attack or decay of a sound waveform, are examples of 'features' which involve relationships between not just two, but many tens or even hundreds of pixels or time points.

The failure of correlation-based learning is most clearly shown by the filters they produce when trained on stationary ensembles of signals. The filters are typically *global* (see figure 2(*a*, *b*)), sensitive to different spatio- or temporal frequencies, but with non-zero weights extending throughout the filter. They reflect only the amplitude spectrum of the signal and ignore the phase spectrum where most of the suspicious *local* coincidences in natural signals take place. An edge in an image, for example, is a coincidence in the phase

* This paper was presented at the Workshop on Information Theory and the Brain, held at the University of Stirling, UK, on 4–5 September 1995.

0954-898X/96/020261+06$19.50

spectrum, since if we were to Fourier analyse it, we would see many sine waves of different frequencies, all aligned in phase where the edge occurred. Correlation-based methods cannot learn edge-detectors, though they often may seem to be doing so by local-windowing of the learnt Fourier components, turning them into Gabor-like filters (see Daugman (1985) for an analysis of the pertinence of Gabor filters to vision).

To illustrate formally that second-order statistics only carry information about the amplitude spectrum, consider the autocorrelation function of a signal, which contains all its second-order structure. The Fourier transform of this is the power spectrum, which is the square of the amplitude spectrum. Thus the two carry identical information.

To demonstrate intuitively that what we consider as the informative part of a natural signal is captured in the phase spectrum, Fourier transform the signal, remove the phase information, and transform it back to the space or time domain. It will then look or sound like noise, typically with a $1/f$ amplitude spectrum. All the visual or auditory features that our perceptual system thinks of as 'signal' will be lost. On the other hand, if we remove the amplitude information, and preserve the phase, the signal will be distorted but remain recognisable.

This points to a curious paradox: correlation-based learning algorithms are sensitive to exactly the part of natural signals which we regard as least meaningful (amplitude), and ignore the part of the signal which we find most meaningful (phase). To encode the phase of signals, we need an algorithm that is sensitive to higher-order statistics.

The problem with higher-order statistics is that there are an infinite number of them. Deciding which to measure *a priori* would be a difficult task. Looking for horizontal bars in an image, for example, we may decide to estimate the average product of all rows of eight pixels occurring in the image. Workers in the field of blind signal processing (Jutten and Herault 1991, Comon 1994, Haykin 1994) have faced these problems, often truncating their higher-order analyses at fourth-order cumulants. However, in Bell and Sejnowski (1995a, b), we described an approach to blind signal processing which was implicitly sensitive to statistics of *all* orders, up to infinity, without having to estimate any one of them explicitly. We now carry this approach over to the domain of feature-learning.

2. Decorrelation, ICA and independent basis functions

Before describing the algorithm, we introduce briefly various matrices involved in decorrelating inputs. Decorrelation is defined as transforming a zero mean vector $\boldsymbol{x} = [x_1, \ldots, x_N]^{\mathrm{T}}$, with a matrix, $\mathbf{W}$, so that the output, $\boldsymbol{u} = \mathbf{W}\boldsymbol{x}$, has a covariance matrix, $\langle \boldsymbol{u}\boldsymbol{u}^{\mathrm{T}} \rangle$, which is diagonal. Solutions to the equation $\mathbf{W}^{\mathrm{T}}\mathbf{W} = \langle \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} \rangle^{-1}$ are all decorrelating matrices.

One of these is given by principal component analysis (PCA), which we will denote $\mathbf{W}_P$. The rows of $\mathbf{W}_P$ are scaled eigenvectors of the covariance matrix, $\langle \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} \rangle$, of the inputs. PCA generally produces global filters which are ordered according to the amplitude spectrum of the signal.

Another solution is the zero-phase one, $\mathbf{W}_Z = \langle \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} \rangle^{-1/2}$ (conveniently given by the MATLAB `sqrtm` function). The rows of $\mathbf{W}_Z$ are local symmetrical filters which are ordered according to the phase spectrum of the signal. These filters flatten the amplitude spectrum, while preserving the phase spectrum of the signal. In other words, they are temporally (or spatially) ordered *whitening* filters, which 'sphere' the data. $\mathbf{W}_Z$ is related to the transforms described by Goodall (1960) and Atick and Redlich (1993).

A third solution is independent component analysis (ICA), or $\mathbf{W}_I$. This matrix not only decorrelates u, but *factorizes* its probability density function, so that $f_{\boldsymbol{u}}(\boldsymbol{u}) = \prod_i f_{u_i}(u_i)$.

This stronger criterion demands zero mutual information between the outputs: $I(u_i, u_j) = 0, \ \forall \ i \neq j$. To achieve ICA we maximise the joint entropy, $H[g(\mathbf{W}x)]$, of the linear transform squashed by a sigmoidal function, $g()$, which is the cumulative density function (c.d.f.) of some 'feature', or signal that we are trying to extract. Often we must guess the c.d.f.'s involved, or in fact there may be no independent solution, so ICA will not always succeed (unlike the above procedures), but it will usually produce something meaningful. Full details and caveats are given in Bell and Sejnowski (1995a, b), where it is reported that good results on super-Gaussian (high kurtosis) feature distributions may be obtained when using the logistic function, $g(u) = (1 + \exp(-u))^{-1}$, in the absence of exact knowledge of particular c.d.f.'s. ICA filters have no particular ordering in phase or frequency and derive their form from both second- and higher-order statistics.

Finally, we describe $\mathbf{W}_B$, the *columns* of which are the 'independent basis functions' of the ensemble of signals (Karhunen *et al* 1995). This may be calculated from the ICA transform by the relation $\mathbf{W}_B = \mathbf{W}_I^{-1}$. The reasoning behind this has been given by Olshausen and Field (1996). Imagine our signals consist of a linear combination of a vector of independently occurring underlying causes, s, each of which is associated with a basis function. Then the signal we receive can be written as $x = \mathbf{W}_B s$. In analogy to the work on source separation , we then recover the independent causes by finding an ICA matrix, $\mathbf{W}_I$, so that $\mathbf{W}_I \mathbf{W}_B = \mathbf{P}$, where $\mathbf{P}$ is a permutation and scaling matrix. When the outputs, u, of our network are the causes, s, scaled and with their order shuffled ($u = \mathbf{P}s$), then the basis functions are just the columns of $\mathbf{W}_I^{-1}$.

To perform the entropy maximizations required for ICA, we update weights incrementally according to the entropy gradient. (For a more leisurely derivation, see Bell and Sejnowski (1995a), and for useful theoretical background, see Nadal and Parga 1995). Defining $y_i = g(u_i)$ to be the sigmoidally transformed output variables, the learning rule is then

$$\Delta \mathbf{W} \propto \frac{\partial H(y)}{\partial \mathbf{W}} = E\left[\frac{\partial \ln |J|}{\partial \mathbf{W}}\right]. \tag{1}$$

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

$$J = \det\left[\frac{\partial y_i}{\partial x_j}\right]_{ij}. \tag{2}$$

In *stochastic* gradient ascent we remove the expected value operator in equation (1), and the derived rule is

$$\Delta \mathbf{W} \propto [\mathbf{W}^{\mathrm{T}}]^{-1} + \hat{y} x^{\mathrm{T}} \tag{3}$$

where $\hat{y} = [\hat{y}_1, \ldots, \hat{y}_N]^{\mathrm{T}}$, the elements of which are

$$\hat{y}_i = \frac{\partial}{\partial y_i}\frac{\partial y_i}{\partial u_i}. \tag{4}$$

In more recent work, we use the natural gradient modification proposed by Amari *et al* (1996), in which the entropy gradient is scaled by the positive definite matrix $\mathbf{W}^{\mathrm{T}}\mathbf{W}$. This speeds convergence and gives the rule (cf equation (3)):

$$\Delta \mathbf{W} \propto \frac{\partial \ln |J|}{\partial \mathbf{W}} \mathbf{W}^{\mathrm{T}}\mathbf{W} = (\mathbf{I} + \hat{y} u^{\mathrm{T}})\mathbf{W}. \tag{5}$$

3. An experiment on tooth-tapping

As a simple test of these principles, we chose a simple musical sound which is local in both frequency in time. The sound is that of a fingernail tapped on a tooth, while the mouth is held in such a position as to resonate at a particular frequency, generating a transient pitch lasting from 5–20 ms. Two such notes (at 700 Hz and 2100 Hz) are illustrated in figure 1(*c*), (*d*).

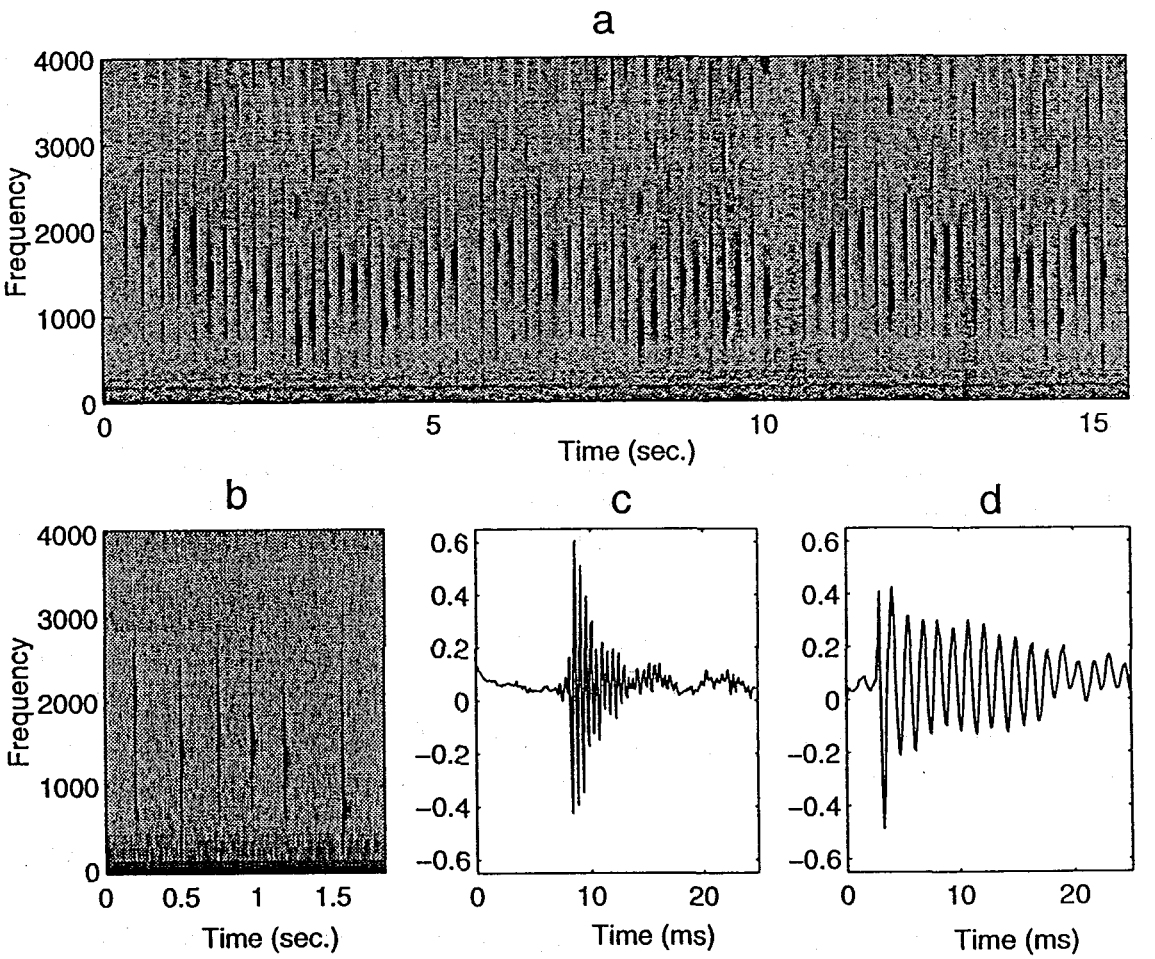

Figure 1. (*a*) Spectrogram of 'Für Elise' by Beethoven played on a tooth. (*b*) Close-up showing six notes, showing how they consist of both a time-local click and a frequency-local tone. (*c*) and (*d*) are, respectively, waveforms of high- and low-frequency tooth taps, the first and last notes in (*b*).

Musical tooth tapping consists of a broadband click (figure 1(*a*), (*b*)), followed immediately by a pure tone with a decaying envelope. This characteristic time structure as well as the intermingling of phase and frequency information in a simple waveform makes it a particularly useful test signal for a system attempting to discover higher-order structure. If our ideas are correct, we should learn filters which reflect both the pitches present and the envelope of the notes. Because the notes are so short, their envelope can be captured in small filters covering only $\frac{1}{80}$ *s*.

3.1. Methods

Fifteen seconds of tooth-tapping of the melody 'Für Elise' by Beethoven were recorded with the microphone of a Sparc-20 workstation at 8 kHz sampling rate (figure 1(*a*)). The mean was subtracted and then 10 000 samples of length 100 ($\frac{1}{80}$ s) were generated from random time-points of the data. The 100×100 covariance matrix of this data was calculated, and from this the PCA ($\mathbf{W}_P$) and zero-phase ($\mathbf{W}_Z$) decorrelating transforms were extracted. The data vectors were pre-whitened ($x \leftarrow \mathbf{W}_Z x$) to speed the subsequent training (see Karhunen *et al* 1995, Bell and Sejnowski 1995b), and an ICA network was trained on the result, using equation (3) with $g(u_i)$ being the logistic function (for which $\hat{y}_i = 1-2y_i$). The weight matrix, $\mathbf{W}$, was initialized to the identity matrix and trained on 24 sweeps through the 10 000 data vectors. The learning rate (per data vector presentation) was dropped from 0.001 to 0.0005 after the first 20 sweeps, in order to anneal the solution. During each sweep, weight changes were accumulated in batches of 50 presentations and then $\mathbf{W}$ was updated, in order to make the vectorized code more efficient. Execution of the 24 learning sweeps took 50 min on a Sparc-20. The full ICA transform was calculated from the result using $\mathbf{W}_I = \mathbf{W}\mathbf{W}_Z$. The independent basis functions were then extracted from the columns of $\mathbf{W}_B = \mathbf{W}_I^{-1}$. All these analyses were performed using MATLAB.

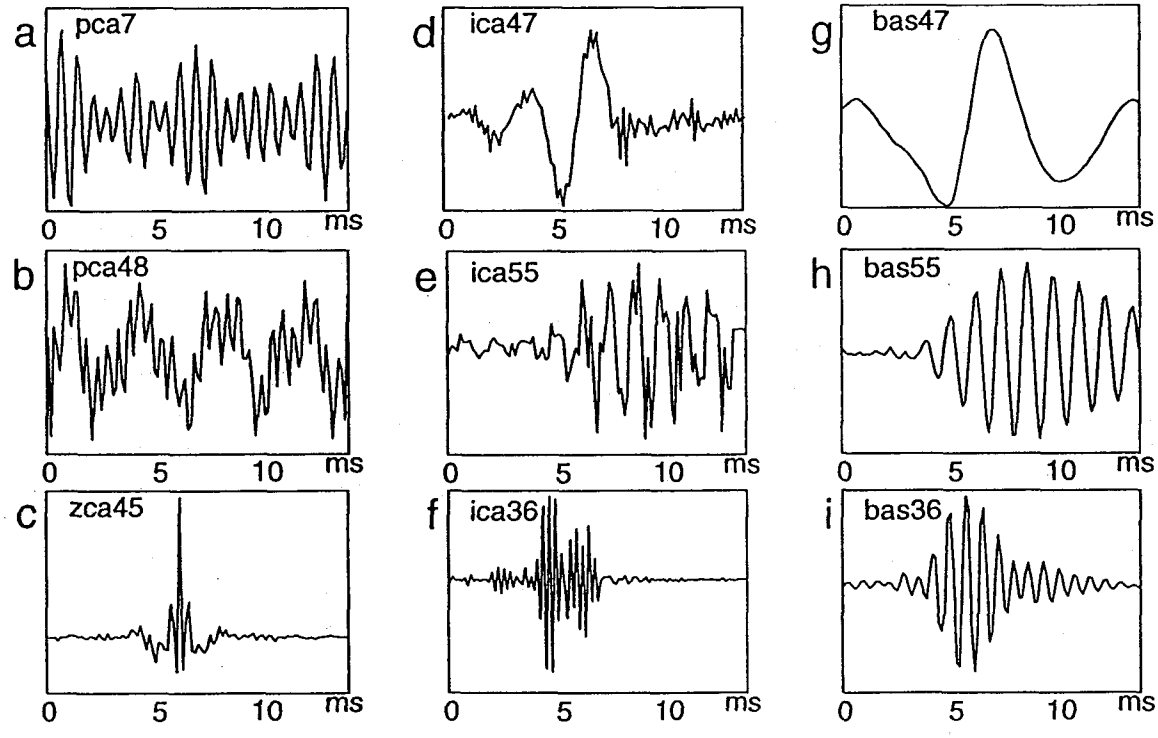

Figure 2. (*a*), (*b*) The seventh and 48th principal components (rows of $\mathbf{W}_P$) of the raw audio data showing only frequency but no phase sensitivity. (*c*) One of the zero-phase decorrelating filters used in prewhitening the data (the 45th row of $\mathbf{W}_Z$). (*d*)–(*f*) three of the filters learnt by the ICA algorithm (3 rows of $\mathbf{W}_I$). (*d*) is sensitive to low-frequency air-conditioning noise, and (*e*) and (*f*) are tooth tap detectors of different frequencies. (*g*)–(*i*) The corresponding three independent basis functions, or waveforms to which these filters optimally respond (three columns of $\mathbf{W}_B$). See text for further explanation.

3.2. Results

Space does not permit the display of all 100 of each filter and basis function, though some results are shown in figure 2. The principal components did not reflect any of the phase information in the signal, as expected. Two of them are shown in figure 2(*a*), (*b*). They are mostly local in frequency, though the deviation we see from this in figure 2(*b*) shows that the training set is not quite stationary. Figure 2(*c*) is an example of one of the $\mathbf{W}_Z$ whitening filters, showing its symmetrical, time-localized appearance (in contrast with the frequency-localized PCA filters). The ICA filters and independent basis functions, three of each of which are shown in figure 2(*d*)–(*f*) and figure 2(*g*)–(*i*) respectively, were all localized in both time and frequency, except for eight low-frequency components. One of the latter is represented in figure 2(*d*), (*g*) and together all eight of them spanned the space of the low-frequency air-conditioning noise visible at the bottom of the spectrograms in figure 1(*a*), (*b*). The other filters and basis functions were all qualitively similar to figure 2(*e*), (*f*) (tooth-tap detectors) and figure 2(*h*), (*i*) (tooth-tap waveforms). Together they covered the frequencies present in the recording, as well as matching well the time course of the tooth-tap waveforms. The one discrepancy is that the sharp attack of the tooth-taps, seen in figure 1(*c*), (*d*), is not reproduced in the basis functions in figure 2(*h*), (*i*). One possible reason is that the broadband click of the attack waveform is statistically independent of the frequency of the note that follows it, so that some of the more broadband highly time-localized basis functions (not shown here) are encoding these portions of the waveform.

4. Discussion

These results show how the unsupervised learning algorithm described in Bell and Sejnowski (1995a) may be used to elucidate the higher-order structure of natural signals, building filters whose associated basis functions represent both the frequency and phase spectrum of the signals. We have also performed experiments on natural images, where the algorithm finds filters and basis functions which have the local, oriented and multiscale properties which the statistics of natural scenes dictate, as argued by Field (1987, 1994). (They are also, of course, what visual cortex provides us with.) Olshausen and Field (1996) have led the search for learning algorithms capable of providing such filters (though see also Intrator 1992). Their scheme, and that of Harpur and Prager (1996) (both of which also appear in this issue) produce qualitatively similar results as ours, though their emphasis is on sparseness rather than statistical independence. Despite this difference, there is every hope that a common perspective will soon be found, given the strong connections between sparse (or minimum entropy) coding on the one hand, and factorial coding on the other (Atick 1992, Barlow 1989, Foldiak 1990, Bell and Sejnowski 1995a, Field 1994).

Acknowledgments

The authors are both with the Howard Hughes Medical Institute in the Computational Neurobiology Laboratory of the Salk Institute. This work was funded the Office of Naval Research and by the Howard Hughes Medical Institute. Many thanks to Bruno Olshausen for showing us how to interpret our results, and to Bruno, David Field, Paul Viola and Nicol Schraudolph for many helpful discussions.

References

Atick J J 1992 Could information theory provide an ecological theory of sensory processing? *Network: Comput. Neural Syst.* **3** 213–51

Atick J J and Redlich A N 1990 Towards a theory of early visual processing *Neural Comput.* **2** 308–20

——1993 Convergent algorithm for sensory receptive field development *Neural Comput.* **5** 45–60

Barlow H B 1989 Unsupervised learning *Neural Comput.* **1** 295–311

Bell A J and Sejnowski T J 1995a An information maximization approach to blind separation and blind deconvolution *Neural Comput.* **7** 1129–59

——1995b Fast blind separation based on information theory, in *Proc. Int. Symp. on Nonlinear Theory and Applications (NOLTA, Las Vegas, December 1995)* **1** (IEICE) pp 43–7

Comon P 1994 Independent component analysis, a new concept? *Signal Process.* **36** 287–314

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 1160–9

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 2370–93

——1994 What is the goal of sensory coding? *Neural Comput.* **6** 559–601

Foldiak P 1990 Forming sparse representations by local anti-Hebbian learning *Biol. Cybern.* **64** 165–70

Goodall M C 1960 Performance of stochastic net *Nature* **185** 557–8

Harpur G F and Prager R W 1996 Development of low entropy coding in a recurrent network *Network: Comput. Neural Syst.* **7** 272–84

Haykin S (ed) 1994 *Blind Deconvolution* (Englewood Cliffs, NJ: Prentice-Hall)

Intrator N 1992 Feature extraction using an unsupervised neural network *Neural Comput.* **4** 98–107

Jutten C and Herault J 1991 Blind separation of sources, part I: an adaptive algorithm based on neuromimetic architecture *Signal Processing* **24** 1–10

Karhunen J, Wang L and Jousensalo J 1995 Neural estimation of basis vectors in independent component analysis *Proc. Int. Conf. on Artificial Neural Networks (ICANN, Paris, 1995)* (Berlin: Springer)

Linsker R 1988 Self-organization in a perceptual network *Computer* **21** 105–17

Miller K D 1988 Correlation-based models of neural development *Neuroscience and Connectionist Theory* ed M Gluck and D Rumelhart (Hillsdale, NJ: Erlbaum) pp 267–353

Nadal J-P and Parga N 1994 Non-linear neurons in the low noise limit: a factorial code maximizes information transfer *Network: Comput. Neural Syst.* **5** 565–81

Oja E 1989 Neural networks, principal components and linear neural networks *Neural Networks* **5** 927–35

Olshausen B A and Field D J 1996 Natural image statistics and efficient coding *Network: Comput. Neural Syst.* **7** 333–9

Sanger T D 1989 Optimal unsupervised learning in a single-layer network *Neural Networks* **2** 459–73