Localized Principal Components of Natural Images - an Analytic Solution *

Yong Liu Harel Shouval
Department of Physics
Institute for Brain and Neural Systems
Box 1843, Brown University
Providence, R. I., 02912
yong@cns.brown.edu
hzs@cns.brown.edu

January 18, 1994

Abstract

The structure of receptive fields in the visual cortex is believed to be shaped by unsupervised learning. A simple variant of unsupervised learning is the extraction of principal components. In this paper, we derived analytically the form of the principal components of natural images. An assumption is made that only small circular regions of the images are being used as training patterns. The derivation relies on results about the correlation function of natural images (Field, 1987). Our results predict both the shapes and the phases of the receptive fields. We also compared these resluts to previous simulation results (Hancock et al., 1992). Finally, the biological relevance of our results is discussed.

1 Introduction

It is generally believed that the shape of the receptive fields in the visual cortex is determined by some form of unsupervised learning. Hebb's postulate (Hebb, 1949) is the cornerstone of most unsupervised learning models in neural networks. The naive Hebbian rule, however, is unstable. This problem is often overcome by adding constraints to this rule. One such constraint is adding a decay term to this rule. It has been proven that a neuron with Hebbian learning rule plus a proper decay term can perform a principal component extraction (Oja, 1982). Furthermore, a neural network with proper lateral inhibition can perform the extraction of several principal components simultaneously (Sanger, 1989; Oja, 1992, eg:).

*to appear in Network: Computation in Neural Systems

There are several computational reasons for extracting principal components. First, in order to compress images, they are the optimal linear projections for minimizing the mean squared reconstruction error (Fukunaga, 1990). Furthermore, it has been argued that a decorrelated representation is important for spotting novel correlations (Barlow, 1989), and principal components neurons indeed eliminate pairwise correlations.

Since the principal components of a set of inputs depend on their covariance matrix, it is reasonable that given this matrix, they can be calculated analytically. A known example is Linsker's multilayered network (Linsker, 1986; Linsker, 1988), which is trained with random noise. The form of the covariance matrix in these models, results from the existence of the Arbor functions between the random input layer and the first layer. This induced correlation was used by several investigators to give an analytic explanation of Linsker's result (Kammen and Yuille, 1988; Yuille et al., 1989; Tang, 1989; MacKay and Miller, 1990a; MacKay and Miller, 1990b)[1].

As has been seen in many experiments (Imbert and Buisseret, 1975; Braastad and Heggelund, 1985, etc.), the plasticity in the visual cortex lasts after the animal opens its eyes. In a normal reared animal, the population of sharply tuned neurons increases monotonically, whereas for dark reared animal it initially increases, but then falls back to the initial level. This implies that the shapes of the receptive fields, in the mature animal, depend on the nature of the environment to which the animals are exposed. Therefore, we believe it is important to model the environment with natural scenes, rather than with artificial environments such as random noise, bars, edges, etc. .

In the model we analize, we assume that the neuron is exposed only to a small circular part of the images. This seems sensible since it restricts the total length of the connections between the image plane and the neuron. This also forces the solutions to be localized. It is important to notice that the results derived here depend on this choice of a circular window. The choice of a circular window, though, seems more biologically plausible than a square window. Furthermore, since the correlation matrix of natural scenes, is nearly radially symmetric, it seems natural to impose a circular boundary, rather than a square boundary which will artificially break this symmetry.

When the input environment is not random, the properties of the correlation function depend on the nature of the environment, and on the preprocessing performed by the visual pathway. However, in this paper, we have simplified the problem by neglecting the effects of this preprocessing, as a first step towards a more complete mathematical understanding of the receptive field structure of neurons in the visual cortex. This approach is different from that undertaken in some of the previous analysis of similar learning rules, (Tang, 1989; MacKay and Miller, 1990a; MacKay and Miller, 1990b, etc.) in that it tires to explain post eye opening plasticity, and therefore uses the correlation function of natural scenes, rather than the artificial correlation functions assumed to be appropriate for explaining prenatal plasticity.

The nature of the correlation function of natural images was investigated by Field (Field, 1987). He found that the spectrum of the correlation function is proportional to the inverse

[1]This model is in principle similar to principal components, although it has some additional terms.

of the square of the frequency. In section 2, we shall use Fields' results to derive analytically the form of the localized principal components. In section 3, we shall show, that under the assumption that the correlation function spectrum has a small non-rotationally symmetric correction, the solutions have a definite phase.

These results partially explain simulation results of Hancock et al. (Hancock et al., 1992), in which the authors used a network of the type proposed by Sanger (Sanger, 1989) to extract principal components from natural scenes. In their simulations, though, they used a smooth Gaussian window, rather than the hard circular window we used, thus altering some of the properties of the resulting receptive fields. When simulation results (Hancock et al., 1992) are compared to our results, we find a good qualitative agreement.

It is clear that the receptive fields obtained here are not identical to the receptive field structures obtained experimentally in the visual cortex. Does this imply that neurons in the first layers of the visual cortex are not principal component analyzers? We shall address this issue in the discussion.

2 The Rotationally Symmetric Solution

The principal components are the eigen-vectors of the covariance matrix. Therefore the equation we try to solve is the eigenvalue problem, i.e., the eigen-equation, which has the form

$$\sum_j C_{ij} W_j = \lambda W_i, \tag{1}$$

where W_i are eigen-vectors, λ is the eigenvalue, and C_{ij} is the covariance matrix which is defined as $C_{ij} = \mathrm{E}[(I_i - \mathrm{E}[I_i])(I_j - \mathrm{E}[I_j])]$ for input pattern $\{I_i\}$. Since we are dealing with two dimensional space, the index i really denotes a point in the two dimensional space, so it is more convenient to rewrite the covariance matrix in the form $C(\mathbf{r}_i, \mathbf{r}'_j)$, which we shall call correlation function. Due to translational invariance, $C(\mathbf{r}_i, \mathbf{r}'_j) = C(\mathbf{r}_i - \mathbf{r}'_j)$. In the continuous limit, the summation will become an integral over $\mathbf{r}'$, thus the eigen-equation becomes

$$\int C(\mathbf{r} - \mathbf{r}')\psi(\mathbf{r}')d^2\mathbf{r}' = \lambda\psi(\mathbf{r}). \tag{2}$$

in which $\psi(\mathbf{r})$ is the continuous limit of the eigen-vectors W_i.

Since the correlation function is radially symmetric, all solutions to the eigen-equation 2 have the form, $\psi_m(\mathbf{r}) = f_m(r)\exp[im(\theta + \phi)]$.

The Fourier transform (spectrum) of the correlation function has the form, $C(\mathbf{k}) = c/\mathbf{k}^2$ where c is a constant (Field, 1987). Hereafter we shall set $c = 1$ for convenience, thus $C(\mathbf{r})$ satisfies

$$\nabla^2 C(\mathbf{r}) = -\delta(\mathbf{r}), \tag{3}$$

which can be readily proven by taking Fourier transformation on both side of this equation. In principle, a gauge $g(\mathbf{r})$ which satisfies $\nabla^2 g(\mathbf{r}) = 0$, can be added to $C(\mathbf{r})$. But the only such gauge, which is radially symmetric, is a constant (denoted as b). This constant will only affect the nature of the radially symmetric, or $m = 0$, solutions. It is important to note that the solution to the set of equations 2, and 3, depends on the boundary condition imposed on them. Imposing a square boundary results in a different set of solutions from those we derive for a circular boundary.

We have chosen to represent the correlation function and receptive fields in the normalized Bessel-Fourier set $W_m(k_i^m \mathbf{r})$ which is zero on the boundary $r = a$ (Jackson, 1975, etc.). The receptive fields are, therefore, a superposition of these functions, i.e., $\psi(\mathbf{r}) = \sum_i A_{mi} W_m(k_i^m \mathbf{r})$. As shown in detail in appendix A, when the correlation function is represented in this set and combined with the eigen-equation, it yields the following matrix equation,

$$\sum_j [C(l)]_{ij} A_{lj} = \lambda A_{li}. \tag{4}$$

in which,

$$[C(l)]_{jj'} = \sum_i \frac{\varepsilon_{lij}\varepsilon_{lij'}}{2\pi N_0(k_i^0)\,(k_i^0)^2}. + b\delta_{l0} V_j V_{j'} \tag{5}$$

where

$$\begin{aligned}
\varepsilon_{lnj} &= \int J_l(k_n^0 r) e^{il\theta} W_l(k_j^l \mathbf{r}) d^2 r \\
&= \begin{cases} \sqrt{2\pi N_0(k_i^0)}\delta_{jn} & \text{for } l = 0 \\[2ex] \sqrt{\frac{2\pi}{N_l(k_j^l)}} \left(\frac{k_j^l}{(k_n^0)^2 - (k_j^l)^2}\right) J_l(k_n^0) J_{l+1}(k_j^l) & \text{for } l > 0 \end{cases}
\end{aligned}$$

and

$$V_j = \frac{2\sqrt{\pi}}{k_j^0}.$$

and N_m is the normalization constant, which takes the value $N_m(k) = (\frac{1}{2})[J_{m+1}(k)]^2$.

This matrix $[C(l)]$ is infinite and in general non-diagonal. However, if $b = 0$, then for the case $l = 0$, it takes the particularly simple diagonal form, $[C(0)]_{ij} = \delta_{ij}/(k_i^0)^2$. Plugging this into the matrix equation, we obtain the radially symmetric solutions to the eigen-equation. They are simply the Bessel-Fourier functions $W_0(k_l^0 \mathbf{r})$ with corresponding eigenvalues $\lambda_{0l} = 1/(k_l^0)^2$.

For $l \neq 0$, it is still possible to obtain a good approximations for the first eigen-vectors, due to the diminishing nature of the eigenvalues $[1/k_i^0]$ of the Bessel equation, the infinite summation involved in computing $C(l)$ can be well approximated by a finite summation. Furthermore, this matrix will have a near-diagonal form. Thus the effective number of Bessel functions of which each eigen-function is composed, at least for small l, will not be large. We use this approximation and solve equation 4 numerically.

Some of these eigen-functions, with the arbitrary choice $b = 0$, are shown in figure 1, in which only the real part ($\propto cos(m\theta)$) has been displayed. In general, the complete

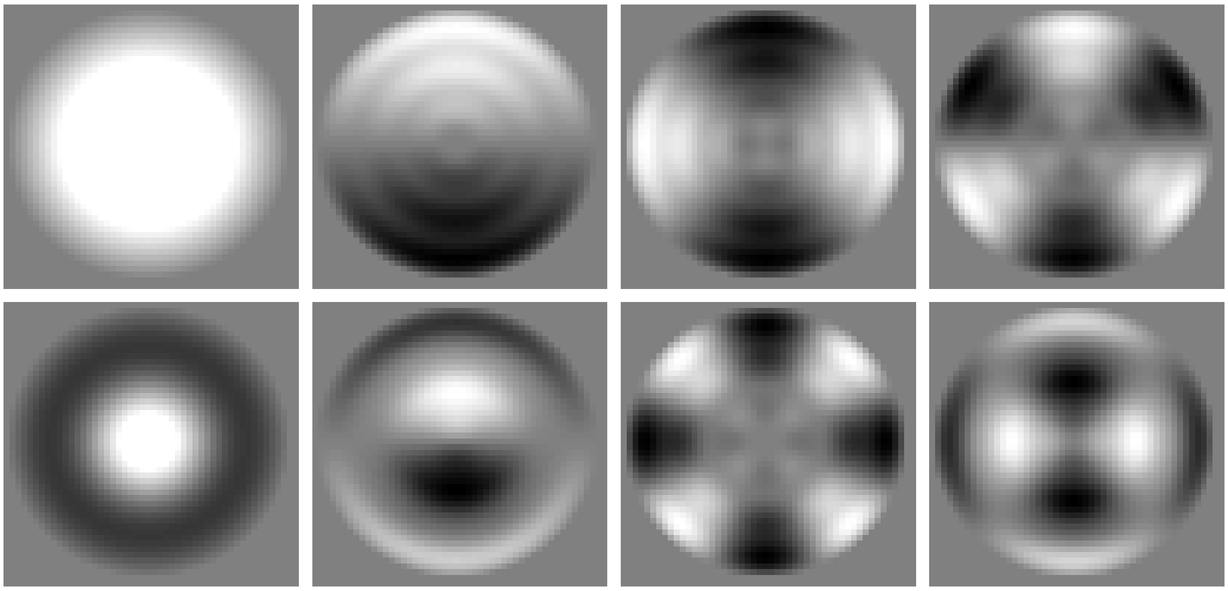

Figure 1: *The shapes of the first eight principal components ($b = 0$ $a = 1$ and $\phi_{mi} = 0$), ordered from left to right, top to bottom.*

set of solutions should include both parts, i.e., $\psi^1_m(\mathbf{r}k^m_i) \propto cos(m(\theta + \phi))$ and $\psi^2_m(\mathbf{r}k^m_i) \propto sin(m(\theta + \phi))$. It should be noted that receptive fields of similar neurons, in different environments with radially symmetric correlation functions(MacKay and Miller, 1990a, eg:), are qualitatively similar since they have the same angular component, they differ in their radial components, and their eigenvalues.

3 Retrieving the Phase

For the solutions above, $\psi^1_{mi}(\mathbf{r}) = \psi^1_m(\mathbf{r}k^m_i) \propto cos(m(\theta + \phi))$ and $\psi^2_{mi}(\mathbf{r}) = \psi^2_m(\mathbf{r}k^m_i) \propto sin(m(\theta + \phi))$, not only do they have undetermined phases ϕ, but also are degenerate, i.e., both have the same eigenvalues. In this section we shall show that if the correlation function is not completely radially symmetric, this degeneracy will be broken, and that the value of the phase ϕ will be determined. The assumption, that the correlation function is not radially symmetric, is not arbitrary, since an inspection of Fields results(Field, 1987, figure 7) reveals that this is indeed the case. These results also explain the feature in the simulation results (Hancock et al., 1992), in which the phases seem to always take the value zero, and the ψ^1_{mi} solution has a different eigenvalue from the ψ^2_{mi} solution. Dependence of the receptive field structure on the unisotropy has also been demonstrated by Baddeley and Hancock (Baddeley and Hancock, 1991) for and artificial environment in which the unisotropy can be systematically varied.

Hereafter, we assume this perturbation term has, in $\mathbf{k}$ space, the form

$$C'(\mathbf{k}) = U(k)T(\theta_{\mathbf{k}}). \tag{6}$$

In order to calculate this perturbation, the representation of this perturbation in the two degenerate eigen-functions $\psi^1_{mi}(\mathbf{r})$ and $\psi^2_{mi}(\mathbf{r})$ has to be calculated. It is easier to perform this in $\mathbf{k}$ space in which the eigen-functions $\psi^1_{mi}(\mathbf{r})$ and $\psi^2_{mi}(\mathbf{r})$ are replaced by their Fourier transformations,

$$\begin{aligned} \psi^1_{mi}(\mathbf{k}) &= f_{mi}(k)cos(m(\theta_{\mathbf{k}} + \phi_{mi})) \\ \psi^2_{mi}(\mathbf{k}) &= f_{mi}(k)sin(m(\theta_{\mathbf{k}} + \phi_{mi})) \end{aligned} \tag{7}$$

in which $f_{mi}(k)$, is the transformation of the radial component of the solutions. If we denote

$$T(\theta_{\mathbf{k}}) = \sum_l t_l cos(l(\theta_{\mathbf{k}} - \alpha_l)) \tag{8}$$

which is the Fourier expansion of $T(\theta_{\mathbf{k}})$. The representation of the perturbation matrix with respect to the two degenerate eigen-functions has the form

$$\begin{aligned} (C'_{(\mu,m,i|\gamma,m,i)})_{(\mu=1,2|\gamma=1,2)} &= (\int \psi^{\mu}_{mi}(\mathbf{k})^* C'(\mathbf{k}) \psi^{\gamma}_{mi}(\mathbf{k}) d^2\mathbf{k})_{(\mu=1,2|\gamma=1,2)} \\ &= g_{mi} \begin{pmatrix} cos(\delta) & sin(\delta) \\ sin(\delta) & -cos(\delta) \end{pmatrix} \end{aligned} \tag{9}$$

in which $\delta = 2\phi_{mi} + 2m\alpha_{2m}$ and $g_{mi} = \frac{\pi}{2} t_{2m} \int U(k)|f_{mi}(k)|^2 kdk$. Since the two eigen-functions are degenerate, any linear combination of these two eigen-functions is an eigen-function of C. Therefore, all we have to do is to find a linear combination of them which diagonalizes the perturbation matrix, i.e., to find the eigenvalues and eigen-vectors of the matrix in equation 9, which are

$$\begin{aligned} &\begin{pmatrix} cos(\delta/2) \\ sin(\delta/2) \end{pmatrix} \\ &\begin{pmatrix} -sin(\delta/2) \\ cos(\delta/2) \end{pmatrix} \end{aligned} \tag{10}$$

with eigenvalues g_{mi} and $-g_{mi}$, respectively.

Thus the eigen-functions and eigenvalues after the perturbation can be readily written out as

$$\begin{aligned} \psi^+_{mi}(\mathbf{k}) &= f_{mi}(r)cos(m(\theta - \alpha_{2m})) \\ \psi^-_{mi}(\mathbf{k}) &= f_{mi}(r)sin(m(\theta - \alpha_{2m})) \end{aligned} \tag{11}$$

with eigenvalues $\lambda^+_{mi} = \lambda_{mi} + g_{mi}$, and $\lambda^-_{mi} = \lambda_{mi} - g_{mi}$, respectively. So the degeneracy is broken. This is in agreement with Hancock's simulations. These solutions have an important feature, i.e., their phases are determined by the properties of the real world correlation function. If the correlation function has a definite symmetry with an inclination angle α, then the solutions would also have the same symmetry angle. Because in this case $\alpha_{2m} = \alpha$ for all m. The spectrums of the correlation function, shown in figure 7 of Field's paper, indeed indicates a symmetry axis along $\alpha = 0$. Thus equation 11 predicts the zero phase result found in Hancock's simulation. When Hancock used images which were tilted by 45 degrees before being scanned, the preferred axis of the receptive fields was found to be 45 degrees. Again this is predicted by equation 11, because the symmetry axis of the correlation function spectrum also gets rotated by 45 degrees due to the rotated images, i.e., $\alpha = 45°$, and thus the solutions also get rotated by 45 degrees.

4 Discussion

We have derived analytically the forms of the localized principal components of natural images, based on the result about the correlation function of natural scenes (Field, 1987). These receptive fields are shown in figure 1. We have also shown that a non-rotationally symmetric perturbation can break the degeneracy and give a definite phase which only depends on the properties of the real world correlation function.

This technique can be generalized to environments with other correlation functions. We are currently using this method for deriving the localized principle components of natural scenes, after they have been passed through a center-surround filter, which corresponds to a retinal preprocessing.

When comparing the receptive fields predicted here, with the receptive fields of neurons in the visual cortex, two major differences stand out. The first is that some of the receptive fields predicted, such as the radially symmetric ones, do not exist in the visual cortex. This difference may be due, at least in part, to the absence of a retinal preprocessing from the model we have analyzed. Another difference, is that for each of the non-radially symmetric receptive fields predicted, only two orientations are necessary to form a complete set. This feature is a property of any method which extracts complete orthogonal sets, and is clearly in contradiction to the near-continuum of orientations found biologically.

A Appendix: Representing the correlation matrix in a Bessel function Base

Using the addition theorem for Bessel functions, it is easy to show that a radially symmetric form for the δ function, is $\delta(\mathbf{r}-\mathbf{r}') = \sum_{il}[1/2\pi N_0(k^0_j)] J^*_l(k^0_j r) J_l(k^0_j r') e^{il(\theta-\theta')}$, where J_l is the lth order Bessel function, k^0_i is the ith zero of J_0, r and θ are the polar coordinates of $\mathbf{r}$, and $N_0(k^0_j)$ is the normalization constant of $J_0(k^0_j \mathbf{r})$.

Therefore, in the same basis, a solution for the correlation function is,

$$C(\mathbf{r}-\mathbf{r}') = \sum_{jl} \frac{J_l^*(k_j^0 r) J_l(k_j^0 r')}{2\pi (k_j^0)^2 \ N_0(k_j^0)} e^{il(\theta-\theta')}.$$

since it solves equation 3.

We shall rewrite this correlation function in terms of the normalized Bessel-Fourier basis W_{mi}. These functions are zero on the boundary and take the form (Jackson, 1975, etc.)

$$W_m(k\mathbf{r}) = \begin{cases} \frac{J_m(k\mathbf{r})}{\sqrt{2\pi N_m(k)}} e^{im\theta} & \text{for } r \leq a \\ 0 & \text{for } r > a \end{cases}$$

In which $N_m(k)$ is the normalization constant, and is equal to

$$N_m(k) = \frac{1}{2}[J_{m+1}(k)]^2.$$

where we have used $a = 1$ for convenience. These W functions solve the differential equation,

$$\nabla^2 W_m = -k^2 W_m. \tag{12}$$

Since these functions are set to be zero on a circular boundary of radius a, thus the solutions for k are quantized, which take discrete values k_i^m where i denotes the ith zero of the function.

Since we can rewrite $J_l(k_n^0 r)e^{il\theta} = \sum_j \varepsilon_{lnj} W_l(k_j^l \mathbf{r})$, where the value of the coefficients is

$$\begin{aligned} \varepsilon_{lnj} &= \int J_l(k_n^0 r) e^{il\theta} W_l(k_j^l \mathbf{r}) d^2 r \\ &= \begin{cases} \sqrt{2\pi N_0(k_i^0)}\,\delta_{jn} & \text{for } l = 0 \\ \\ \sqrt{\frac{2\pi}{N_l(k_j^l)}} \left(\frac{k_j^l}{(k_n^0)^2-(k_j^l)^2}\right) J_l(k_n^0) J_{l+1}(k_j^l) & \text{for } l > 0 \end{cases} \end{aligned}$$

we can obtain a representation of the correlation function in terms of the Bessel-Fourier set,

$$C(\mathbf{r}-\mathbf{r}') = \sum_{ljj'} [C(l)]_{jj'} W_l(k_j^l \mathbf{r}) W_l^*(k_{j'}^l \mathbf{r}'),$$

where

$$[C(l)]_{jj'} = \sum_i \frac{\varepsilon_{lij}\varepsilon_{lij'}}{2\pi N_0(k_i^0)\ (k_i^0)^2}.$$

This representation of $C(\mathbf{r}-\mathbf{r}')$ is unique only up to a constant b, which adds the term $\delta_{l0} V_j V_{j'}$ to the representation of the correlation matrix, where $V_j = \int W_0(k_j^0 \mathbf{r}) d^2 r = 2\sqrt{\pi}/k_j^0$.

Thus plugging this form of the correlation function into the eigen-equation 2, representing the eigen-functions as a sum of this complete set, or, $\psi(\mathbf{r}) = \sum_{mn} A_{mn} W_m(k_n^m \mathbf{r})$, and using the orthogonality relations, we obtain the matrix equation,

$$\sum_j [C(l)]_{ij} A_{lj} = \lambda A_{li}.$$

B Acknowledgments

The authors thank Professor David Gottlieb, all the members of Institute for Brain and Neural Systems, in particular, professor Leon N Cooper, Charlie Law and Michael P. Perrone. We would also like to thank Tal Grossman and Zili Liu for helpful comments. Y. L. also thanks Cong Zhang for her love and encouragement. This research was supported by grants from ONR, NSF and ARO.

References

Baddeley, R. and Hancock, P. (1991). A statistical analysis of natural images matchs psychophysically derived orientation tuning curves. *Proc Roy Soc B*, 246(17):219–223.

Barlow, H. B. (1989). Unsupervised learning. *Neural Computation*, 1:295–311.

Braastad, B. O. and Heggelund, P. (1985). Development of spatial receptive-field organization and orientation selectivity in kitten's visual cortex. *J. Neurophysiol.*, 42:1158–1178.

Field, D. J. (1987). Relations between the statistics of natural images and the response properities of cortical cells. *J. Opt. Soc. Am. A*, 4(12):2379–2394.

Fukunaga, K. (1990). *Introduction to Statitical Pattern Recognition (2rd Ed.).* Academic Press, Inc.

Hancock, P. J., Baddeley, R. J., and Smith, L. S. (1992). The principal components of natural images. *Network*, 3:61–70.

Hebb, D. O. (1949). *The Organization of Behavior.* Wiley.

Imbert, M. and Buisseret, P. (1975). Receptive field characteristics and plastic properties of visual cortical cells in kittens reared with or without visual experience. *Exp. Brain Res.*, 22:25–36.

Jackson, J. D. (1975). *Classical Electrodynamics.* Wiley.

Kammen, D. M. and Yuille, A. L. (1988). Spontaneous symmetry-breaking energy function and the emergence of selective cortical cells. *Bio. Cybern.*, 59:23–31.

Linsker, R. (1986). From basic network principles to neural architecture. *Proc. Natl. Acad. Sci. USA*, 83:7508–12, 8390–4, 8779–83.

Linsker, R. (1988). Self-organization in a perceptual network. *Computer*, pages 105–17.

MacKay, D. J. and Miller, K. D. (1990a). Analysis of Linsker's simulation of Hebbian rules. *Neural Computation*, 1:173–187.

MacKay, D. J. and Miller, K. D. (1990b). Analysis of Linsker's simulations of Hebbian rules to linear networks. *Network*, 1:257–297.

Oja, E. (1982). A simplified neuron model as a principal component analyzer. *J. Math. Bio.*, 15:267–273.

Oja, E. (1992). Principal components, minor components, and linear neural networks. *Neural Networks*, 5:927–935.

Sanger, T. D. (1989). Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural Networks*, 2:459–473.

Tang, D. S. (1989). Information-theoretic solutions to early visual information processing: Analytic results. *Phys. Rev. A*, 40(11):6626–35.

Yuille, A. L., Kammen, D. M., and Cohen, D. S. (1989). Quadrature and the development of orientation selective cells by hebb rules. *Bio. Cybern.*, 61:183–194.