

---

# Receptive field formation in natural scene environments: comparison of kurtosis, skewness, and the quadratic form of BCM

---

Brian S. Blais<sup>\*</sup> N. Intrator<sup>\*</sup> H. Shouval Leon N Cooper  
Brown University Physics Department and  
Institute for Brain and Neural Systems  
Brown University  
Providence, RI 02912  
{bblais,nin,inc}@cns.brown.edu

## Abstract

We study several statistically and biologically motivated learning rules using the same visual environment and neuronal architecture. This allows us to concentrate on the feature extraction and neuronal coding properties of these rules. We find that the quadratic form of the BCM rule behaves in a manner similar to a kurtosis maximization rule when the distribution contains kurtotic directions, although the BCM modification equations are computationally simpler.

## 1 Introduction

BCM synaptic modification functions (Bienenstock et al., 1982) are characterized by a negative region for small postsynaptic depolarization, a positive region for large postsynaptic depolarization, and a threshold which moves so as to stabilize the learning. In what follows we investigate several specific modification functions that have these general properties and study their feature extraction properties in a natural scene environment. Several of these, such as skewness and kurtosis, are standard statistical measures (Kendall and Stuart, 1977) based on polynomial moments. We compare these with the quadratic form of BCM (Intrator and Cooper, 1992). By subjecting all of the learning rules to the same input statistics and retina/LGN preprocessing and by studying in detail the single neuron case, we eliminate possible network/lateral interaction effects and can examine the properties of the learning rules themselves.

## 2 Properties of natural scene statistics

It is known that natural images produce long-tailed distributions (Daugman, 1988; Field, 1994). It has further been argued that local linear transformations such as Gabor filters or center-surround produce exponential-tailed histogram (Ruderman, 1994). Reasons for that vary from the specific arrangements of the Fourier phases of natural images (Field, 1994) or the existence of edges. As exponential distribution is optimal from the view point of information theory under the assumption of positive and fixed average activity (Ruderman, 1994; Levy and Baxter, 1996; Intrator, 1996), it is a natural candidate for detailed study in conjunction with neuronal learning rules.

---

<sup>\*</sup>Corresponding author. Current address: Box 1843, Brown University, Providence RI 02912

<sup>\*</sup>On leave, School of Mathematical Sciences, Tel-Aviv University.

<sup>o</sup>This work supported by the Charles A. Dana Foundation, the Office of Naval Research, and the National Science Foundation

### 3 Exploratory projection pursuit and feature extraction

Projection pursuit (PP) methods seek features which emphasize the non-Gaussian nature of distributions (Huber, 1985, for review). They seek structure that is exhibited by (semi) linear projections of the data. The relevance to neural network theory is clear, since the activity of a neuron is largely believed to be a semi linear function of the projection of the inputs on the vector of synaptic weights. Diaconis and Freedman (1984) show that for most high-dimensional clouds (of points), most low-dimensional projections are approximately Gaussian. This finding suggests that important information in the data is conveyed in those directions whose single dimensional projected distribution is far from Gaussian. Intrator (1990) has shown that a BCM neuron can find structure in the input distribution that exhibits deviation from Gaussian distribution in the form of multi-modality in the projected distributions. Since clusters can not be found directly in the data due to its sparsity, this type of deviation, which is measured by the first three moments of the distribution, is particularly useful for finding clusters in high dimensional data and is thus useful for classification or recognition tasks.

In this paper, we want to gain some insight on the statistics of small patches of natural scenes through their projections and study various variants of polynomial-moment projection indices. The most common measures for deviation from Gaussian distribution are skewness and kurtosis which are functions of the first four moments of the distribution. Rules based on these statistical measures satisfy the BCM conditions proposed in Bienenstock et al. (1982), including a threshold-based stabilization, but the details, and some of the qualitative features of the stabilization are different. Some of these differences are seemingly important, while others seem not to affect the results significantly. For comparison, we present results for the quadratic form of the BCM modification equation, though one should note that this is not the only form which could be used. In addition, there are some learning rules, such as the ICA rule of Bell and Sejnowski (1997), which have been used with natural scene inputs to produce oriented receptive fields but are not in this study. We omit these because they do not have a straightforward single cell implementation, and would thus detract from our goal of comparing rules with the same input structure and neuronal architecture. For a related study discussing projection pursuit and BCM see (Press and Lee, 1996).

As we are interested in positive neuronal activities only, we denote by  $c$  the rectified activity  $\sigma(\mathbf{d} \cdot \mathbf{m})$  and assume that the sigmoid is a smooth monotone function with a positive output (a slight negative output is also allowed).  $\sigma'$  denotes the derivative of the sigmoidal. The rectification is required for all rules that depend on odd moments because these vanish in a symmetric distribution such as natural scenes. We also demonstrate later that the rectification makes little difference on learning rules that depend on even moments.

**Skewness 1** This measures the deviation from symmetry (Kendall and Stuart, 1977, for review) and is of the form:

$$S_1 = E[c^3]/E^{1.5}[c^2]. \quad (1)$$

A maximization of this measure via gradient ascent gives

$$\nabla S_1 = \frac{1}{\Theta_M^{1.5}} E [c (c - E[c^3]/E[c^2]) \sigma' \mathbf{d}], \quad (2)$$

where  $\Theta_m$  is defined as  $E[c^2]$ .

**Skewness 2** A similar measure which requires some stabilization mechanism is given by

$$S_2 = E[c^3] - E^{1.5}[c^2]. \quad (3)$$

This measure has a gradient of the form

$$\nabla S_2 = 3E [c^2 - c\sqrt{E[c^2]}] = 3E [c (c - \sqrt{\Theta_M}) \sigma' \mathbf{d}], \quad (4)$$

**Kurtosis 1** Kurtosis measures deviation from Gaussian distribution mainly in the tails of the distribution. It has the form

$$K_1 = E[c^4]/E^2[c^2] - 3. \quad (5)$$

This measure has a gradient of the form

$$\nabla K_1 = \frac{1}{\Theta_M^2} E [c (c^2 - E[c^4]/E[c^2]) \sigma' \mathbf{d}]. \quad (6)$$

**Kurtosis 2** As before, there is a similar form which requires some stabilization:

$$K_2 = E[c^4] - 3E^2[c^2]. \quad (7)$$

This measure has a gradient of the form

$$\nabla K_2 = 4E [c^3 - cE[c^2]] = 3E [c(c^2 - \Theta_M)]\sigma' \mathbf{d}. \quad (8)$$

In all the above, the maximization of the measure can be used as a goal for projection seeking, so the variable  $c$  can be thought of as a (nonlinear) projection of the input distribution onto a certain vector of weights, and the maximization then defines a learning rule for this vector of weights. Under this framework, it is easy to stabilize the above learning rules by requiring for example that the vector of weights, which we denote by  $m$ , has a fixed norm, say  $\|m\|=1$ . The multiplicative forms of both kurtosis and skewness do not require this type of stabilization, due to the normalizing factor  $1/\Theta_M^p$  in each rule.

**Quadratic BCM** The Quadratic BCM (QBCM) measure as given in (Intrator and Cooper, 1992) is of the form

$$\text{QBCM} = \frac{1}{3}E[c^3] - \frac{1}{4}E^2[c^2]. \quad (9)$$

Maximizing this form using gradient ascent gives the learning rule:

$$\nabla \text{QBCM} = E [c^2 - cE[c^2]] = E[c(c - \Theta_M)]\sigma' \mathbf{d}. \quad (10)$$

Unlike the above, the Quadratic BCM rule does not require any additional stabilization. This turns out to be an important property, since additional information can then be transmitted using the resulting norm of the weight vector  $m$  (Intrator, 1996).

It is important to note that the Quadratic BCM rule is only one of many possible forms for BCM modification. In fact, both skewness measures clearly follow the same criteria initially proposed by BCM and thus can be seen as statistically motivated variants of BCM. The kurtosis measures use a different form of stabilization, and therefore cannot be so easily identified as BCM-like, but they do share many of the general properties of BCM modification.

## 4 Methods

In order to study the properties of the different learning rules in a natural scene environment, we use patches from 12 images of natural scenes as input to single neurons. To help understand the behavior of the learning under the different rules, we look at the effects of two different types of preprocessing of the images for each of the learning rules. The first is a Difference of Gaussians (DOG) filter, which is commonly used to model the processing done in the retina (Law and Cooper, 1994). The second is a whitening filter, used to eliminate the second order correlations (Oja, 1995; Bell and Sejnowski, 1995). Whitening the data in this way allows one to use learning rules which are dependent on higher moments of the data, but are particularly sensitive to the second moment.

At each iteration of the learning, a 13 by 13 patch is taken from the preprocessed (either DOGED or whitened) images and presented to the neuron. The moments of the output,  $c$ , are calculated iteratively using

$$E[c^n(t)] = \frac{1}{\tau} \int_{-\infty}^t c^n(t') e^{-(t-t')/\tau} dt'$$

In the cases where the learning rule is underconstrained (ie.  $K_2$  and  $S_2$ ) we also normalize the weights each iteration.

## 5 Results

### 5.1 Receptive Fields

The resulting receptive fields formed are shown in Figure 1 for both the DOGED and whitened images. Every learning rule developed oriented receptive fields. The multiplicative versions of kurtosis and skew, as well as Quadratic BCM, sampled from many orientations. The rules minimizing the additive forms of kurtosis and skew, however, did not. These rules seem to have a strong dependence on the second moment, as seen both by the resemblance of the receptive fields to those obtained from PCA (Shouval and Liu, 1996) and also the fact that the problem disappears with the whitened input. The multiplicative skewness rule gives receptive fields with larger spatial frequencies than either Quadratic BCM or the multiplicative kurtosis rule. This also disappears with the whitened inputs.

Receptive Fields from Natural Scene Input

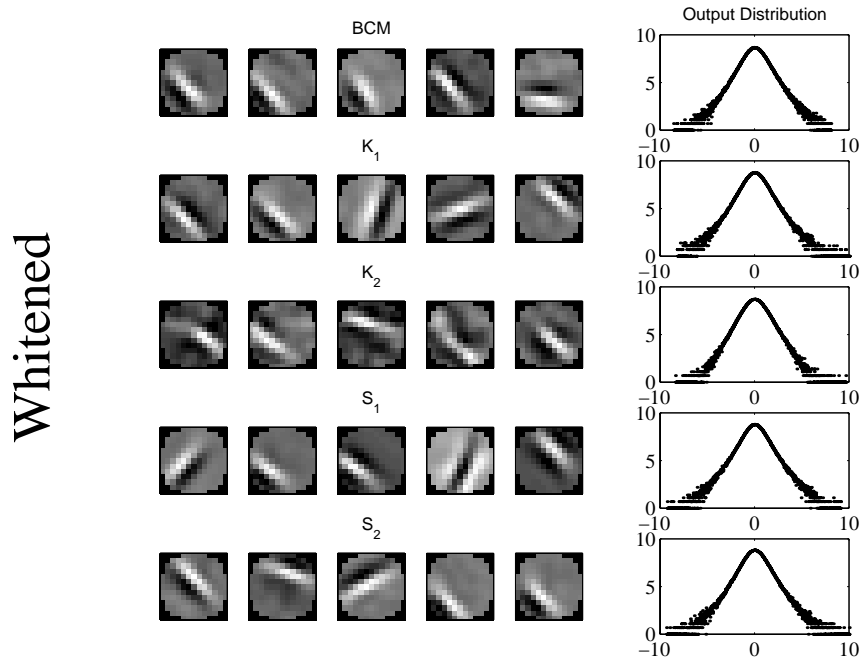
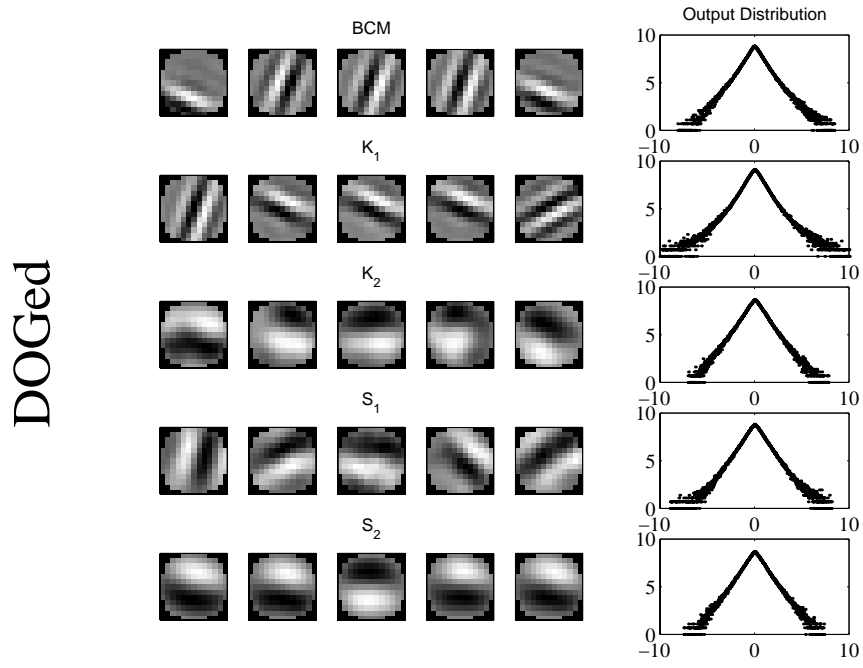


Figure 1: Receptive fields using DOGed image input (upper graphs) and whitened image input (lower graphs), obtained from learning rules maximizing (from top to bottom) the Quadratic BCM objective function, Kurtosis (multiplicative), Kurtosis (additive), Skew (multiplicative), and Skewness (additive). Shown are five examples (left to right) from each learning rule as well as the normalized output distribution, before the application of the rectifying sigmoid.

An objection may be made that the receptive fields formed are caused almost entirely by the application of the rectifying sigmoid. This sigmoid is not needed for rules dependent only on the even powered moments,

such as kurtosis. Figure 2 demonstrates both that the removal of the sigmoid and the removal of the mean from the moments calculations does not substantially affect the resulting receptive fields of the kurtosis rules.

### Exploring Modifications to the Kurtosis rules

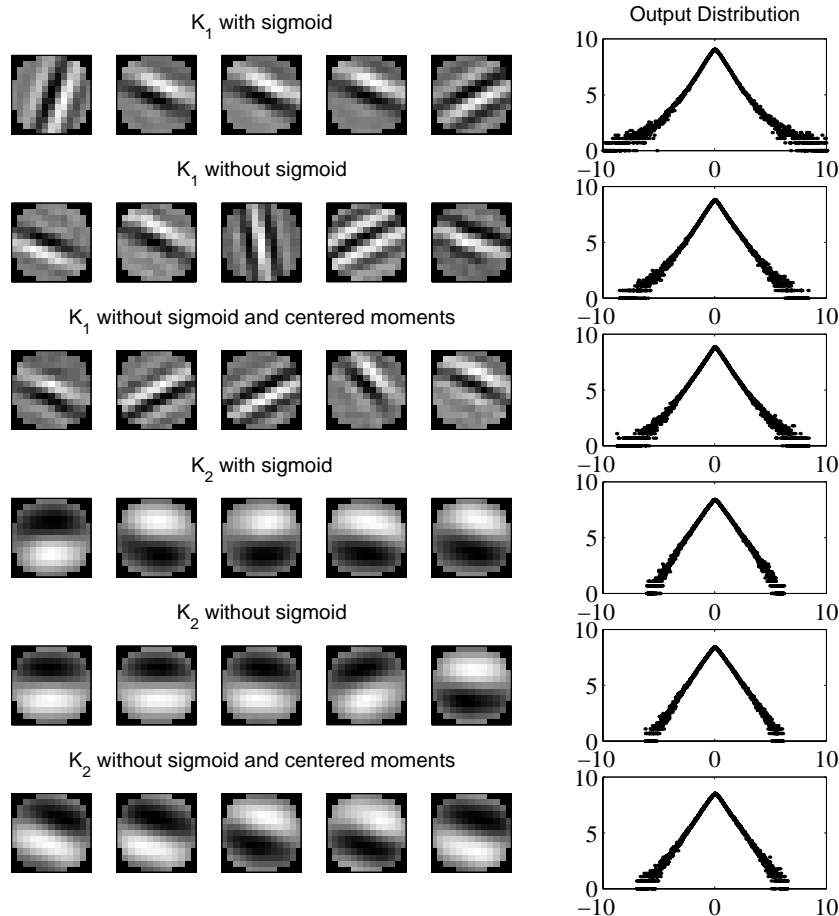


Figure 2: Receptive fields using DOGED image input, obtained from learning rules maximizing (from top to bottom) multiplicative form kurtosis with rectified outputs, non-rectified outputs, and non-rectified outputs with centered moments respectively, and additive form kurtosis with rectified outputs, non-rectified outputs, and non-rectified outputs with centered moments respectively. Shown are five examples (left to right) from each learning rule and the corresponding output distribution.

Note also that the choice of 13 by 13 receptive fields was made only for computational efficiency. Figure 3 shows some 21 by 21 receptive fields and it is clear that little difference is made.

## 5.2 Structure Removal: Sensitivity to Outliers

Learning rules which are dependent on large polynomial moments, such as Quadratic BCM and kurtosis, tend to be sensitive to the tails of the distribution. This property implies that neurons are highly responsive, and sensitive, to the outliers, and consequently leads to a sparse coding of the input signal. We would like to address the degree to which the neurons form a sparse code, by directly testing how much of the input distribution is required to maintain the RF. This can be done in a straightforward and systematic fashion.

First, the neuron is trained in the visual environment until the oriented RF forms properly. Next, those patterns which yield responses in the tail of the distribution are removed. The number of input patterns needed to be removed in order to cause a change in the receptive field gives a direct measure of the sparsity of the coding. Finally, the training is continued using the reduced input environment. This process can continue recursively, sequentially removing more structure from the input environment. The results of this are shown in Figure 4. For Quadratic BCM and kurtosis, one need only delete one percent of the input patterns to change the receptive field, which suggests that the neuron is indeed coding the information in a very sparse manner. For the rule for

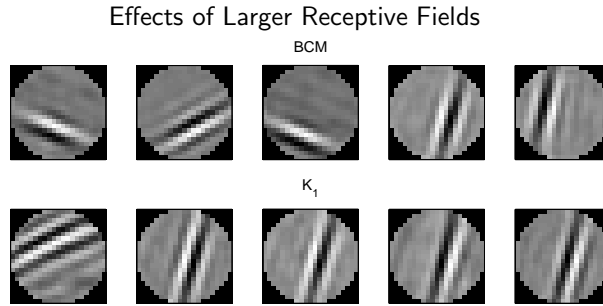


Figure 3: Large receptive fields using DOGED image input, obtained from the Quadratic BCM learning rule and the rule maximizing the multiplicative form of kurtosis.

maximizing skew, more than five percent are needed to alter the receptive field properties.

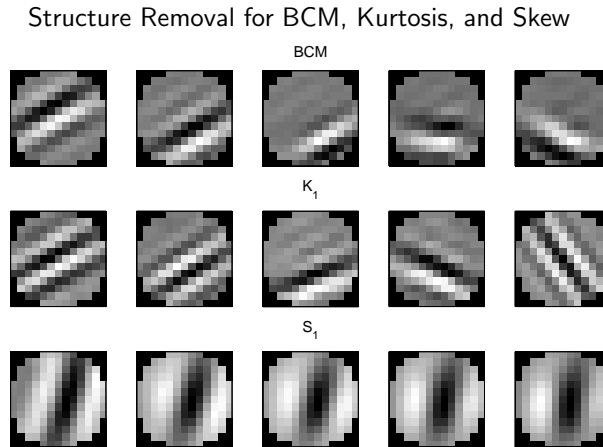


Figure 4: Receptive fields resulting from structure removal using the Quadratic BCM rule, the rule maximizing the multiplicative form of kurtosis and skewness. The RF on the far left for each rule was obtained in the normal input environment. The next RF to the right was obtained in a reduced input environment, whose patterns were deleted that yielded the strongest 1% of responses from the RF to the left. This process was continued for each RF from left to right, yielding a final removal of about five percent of the input patterns.

When the small, but important, part of the input distribution is deleted (namely, the tails of the distribution), the neuron seeks a different RF. This occurs in both the BCM and kurtosis learning rules, and most likely occurs in other rules that seek kurtotic projections. It is important to note, however, that patterns must be deleted from *both* sides of the distribution for any rule that does not use the rectifying sigmoid because the strong *negative* responses carry as much structure as the strong positive ones. Such responses are not biologically plausible, so they wouldn't be part of the encoding process in real neurons.

It is also interesting to observe that the RF found after structure removal is initially of the same orientation, but of different spatial phase. Once enough input patterns are removed, the RF becomes oriented in a different direction. If the process is continued, all of the orientations and phases would be obtained.

## 6 Discussion

This study attempts to compare several learning rules which have some statistical or biological motivation, or both. We have used natural scenes to gain some more insight about the statistics underlying natural images. There are several outcomes from this study:

- All rules used found kurtotic distributions. This should not come as a surprise as there are suggestions that a large family of linear filters can find kurtotic distributions (Ruderman, 1994).
- The Quadratic BCM and the multiplicative version of kurtosis are less sensitive to the second moments of the distribution and produce oriented receptive fields even when the data is not sphered. The subtractive

version of kurtosis is sensitive and produces oriented RF only after sphering the data (Friedman, 1987; Field, 1994).

- Both Quadratic BCM and kurtosis are sensitive to the tails of the distribution. This sensitivity is so high that the RF changes due to elimination of the upper 1% portion of the distribution (Figure 4). The change in RF is gradual; at first, removal of some of the inputs results in RFs that have the same orientation but a different phase, once more patterns from the upper portion of the distribution are removed, different RF orientations are found. This finding gives some indication to the kind of inputs the cell is most selective to (values below its highest 99% selectivity), these are inputs with same orientation but with different phase (different locality of RF). The sensitivity to small portions of the distribution represents the other side of the coin of sparse coding. It should be further studied as it may reflect some fundamental instability of kurtotic approaches.
- Rectified skewness rule can also find oriented RF's. Its sensitivity to the upper parts of the distribution is not so dramatic and thus, the RFs do not change much when few percent of the upper distribution are removed.
- Both Quadratic BCM and kurtosis rules have a built in second order normalization (kurtosis via division and BCM via subtraction) so that they are not very sensitive to second-order structure. This is clear from the results about DOG-processed vs. whitened inputs.
- Kurtotic rules can find high kurtosis in either symmetric or rectified distributions. This is not the case for Quadratic BCM rule which requires rectified distributions.
- The bottom line of this study is the fact that the Quadratic BCM learning rule which has been advocated as a projection index for finding multi-modality in high dimensional distribution, performs well under kurtotic distributions as well and can find directions of high kurtosis. We have preliminary indications that the converse is not true, namely, kurtosis measure does not perform well under distribution that are bi- or multi-modal. This will be shown elsewhere.

## References

- Bell, A. J. and Sejnowski, T. J. (1995). An information-maximisation approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159.
- Bell, A. J. and Sejnowski, T. J. (1997). The independent components of natural scenes are edge filters. *Vision Research*. in press.
- Bienenstock, E. L., N Cooper, L., and Munro, P. W. (1982). Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *Journal Neuroscience*, 2:32–48.
- Daugman, J. G. (1988). Complete discrete 2D Gabor transforms by neural networks for image analysis and compression. *IEEE Transactions on ASSP*, 36:1169–1179.
- Diaconis, P. and Freedman, D. (1984). Asymptotics of graphical projection pursuit. *Annals of Statistics*, 12:793–815.
- Field, D. J. (1994). What is the goal of sensory coding. *Neural Computation*, 6:559–601.
- Friedman, J. H. (1987). Exploratory projection pursuit. *Journal of the American Statistical Association*, 82:249–266.
- Huber, P. J. (1985). Projection pursuit. (with discussion). *The Annals of Statistics*, 13:435–475.
- Intrator, N. (1990). A neural network for feature extraction. In Touretzky, D. S. and Lippmann, R. P., editors, *Advances in Neural Information Processing Systems*, volume 2, pages 719–726. Morgan Kaufmann, San Mateo, CA.
- Intrator, N. (1996). Neuronal goals: Efficient coding and coincidence detection. In Amari, S., Xu, L., Chan, L. W., King, I., and Leung, K. S., editors, *Proceedings of ICONIP Hong Kong. Progress in Neural Information Processing*, volume 1, pages 29–34. Springer.
- Intrator, N. and Cooper, L. N. (1992). Objective function formulation of the BCM theory of visual cortical plasticity: Statistical connections, stability conditions. *Neural Networks*, 5:3–17.
- Kendall, M. and Stuart, A. (1977). *The Advanced Theory of Statistics*, volume 1. MacMillan Publishing, New York.
- Law, C. and Cooper, L. (1994). Formation of receptive fields according to the BCM theory in realistic visual environments. *Proceedings National Academy of Sciences*, 91:7797–7801.
- Levy, W. B. and Baxter, R. A. (1996). Energy efficient neural codes. *Neural Computation*, 8:531–543.
- Oja, E. (1995). The nonlinear pca learning rule and signal separation - mathematical analysis. Technical Report A26, Helsinki University, CS and Inf. Sci. Lab.
- Press, W. and Lee, C. W. (1996). Searching for optimal visual codes: Projection pursuit analysis of the statistical structure in natural scenes. In *The Neurobiology of Computation: Proceedings of the fifth CNS conference*. Plenum Publishing Corporation.
- Ruderman, D. L. (1994). The statistics of natural images. *Network*, 5:517–548.
- Shouval, H. and Liu, Y. (1996). Principal component neurons in a realistic visual environment. *Network*, 7:3. In Press.