LETTER Communicated by Erkki Oja

Receptive Field Formation in Natural Scene Environments: Comparison of Single-Cell Learning Rules

Brian S. Blais
N. Intrator
H. Shouval
Leon N. Cooper
Physics Department and Institute for Brain and Neural Systems, Brown University, Providence, RI 02912, U.S.A.

We study several statistically and biologically motivated learning rules using the same visual environment: one made up of natural scenes and the same single-cell neuronal architecture. This allows us to concentrate on the feature extraction and neuronal coding properties of these rules. Included in these rules are kurtosis and skewness maximization, the quadratic form of the Bienenstock-Cooper-Munro (BCM) learning rule, and single-cell independent component analysis. Using a structure removal method, we demonstrate that receptive fields developed using these rules depend on a small portion of the distribution. 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

Recently several learning rules that develop simple cell-like receptive fields in a natural image environment have been proposed (Law & Cooper, 1994; Olshausen & Field, 1996; Bell & Sejnowski, 1997). The details of these rules are different, as is their computational reasoning; however, all depend on statistics of order higher than two, and all produce sparse distributions.

In a sparse distribution, most of the mass of the distribution is concentrated around zero, and the rest of the distribution extends much farther out. In other words, a neuron that has sparse response responds strongly to a small subset of patterns in the input environment and weakly to all others. Bimodal distributions, for which the mode at zero has significantly more mass then the nonzero mode, and exponential distributions are examples of sparse distributions, whereas gaussian and uniform distributions are not considered sparse. It is known that many projections of the distribution of natural images have long-tailed, or exponential, distributions (Daugman, 1988; Field, 1994). It has been argued that local linear transformations such as Gabor filters or center-surround produce exponential-tailed histograms

(Ruderman, 1994). Reasons given vary from the specific arrangements of the Fourier phases of natural images (Field, 1994) to the existence of edges. Since the exponential distribution is optimal from the viewpoint of information theory under the assumption of positive and fixed average activity (Ruderman, 1994; Levy & Baxter, 1996; Intrator, 1996; Baddeley, 1996), it is a natural candidate for detailed study in conjunction with neuronal learning rules.

In what follows we investigate several specific modification functions that have the general properties of BCM synaptic modification functions (Bienenstock, Cooper, & Munro, 1982) and study their feature extraction properties in a natural scene environment. BCM synaptic modification functions are characterized by a negative region for small postsynaptic depolarization, a positive region for large postsynaptic depolarization, and a threshold that moves and switches between the Hebbian and anti-Hebbian regions. Several of the rules we consider are derived from standard statistical measures (Kendall & Stuart, 1977), such as skewness and kurtosis, based on polynomial moments. We compare these with the quadratic form of BCM (Intrator & Cooper, 1992), though this is not the only form that could be used. By subjecting all of the learning rules to the same input statistics and retina/lateral geniculate nucleus (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.

We start with a motivation for the learning rules used in this study and then present the initial results. We then explore some of the similarities and differences between the rules and the receptive fields they form. Finally, we introduce a procedure for directly measuring the sparsity of the representation a neuron learns; this gives us another way to compare the learning rules and a more quantitative measure of the concept of sparse representations.

2 Motivation

We use two methods for motivating the use of the particular rules. One comes from projection pursuit (Friedman, 1987), where we use an energy function to find directions where the projections of the data are nongaussian (Huber, 1985, for review); the other is independent component analysis (ICA) (Comon, 1994), where one seeks directions where the projections are statistically independent. These methods are related, but they provide two different approaches to this work.

2.1 Exploratory Projection Pursuit and Feature Extraction. 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. There is, however, some indication (Zetzsche, 1997), that for natural images,

random local projections yield somewhat longer-tailed distributions than gaussian. We can still justify this approach, because interesting structure can still be found in non-random directions that yield projections farther 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 multimodality in the projected distributions. 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 through the search for multimodality in the projected distribution rather than in the original high-dimensional space. It is thus useful for classification or recognition tasks. In the natural scene environment, however, the structure does not seem to be contained in clusters: projection indices that seek clusters never find them. In this work we show that the BCM neuron can still find interesting structure in nonclustered data.

The most common measures for deviation from gaussian distribution are skewness and kurtosis, which are functions of the first three and four moments of the distribution, respectively. Rules based on these statistical measures satisfy the BCM conditions proposed by Bienenstock et al. (1982), including a threshold-based stabilization. The details of these rules and some of the qualitative features of the stabilization are different, however. Some of these differences are seemingly important, while others seem not to affect the results significantly. In addition, there are some learning rules, such as the ICA rule of Bell and Sejnowski (1997) and the sparse coding algorithm of Olshausen and Field (1996), which have been used with natural scene inputs to produce oriented receptive fields. We do not include these in our comparison because the learning is not based on the activity and weights of a single neuron, and thus detract from our immediate goal of comparing rules with the same input structure and neuronal architecture.

2.2 Independent Component Analysis. Recently it has been claimed that the independent components of natural scenes are the edges found in simple cells (Bell & Sejnowski, 1997). This was achieved through the maximization of the mutual entropy of a set of mixed signals. Others (Hyvarinen & Oja, 1997) have claimed that maximizing kurtosis, with the proper constraints, can also lead to the separation of mixed signals into independent components. This alternate connection between kurtosis and receptive fields leads us into a discussion of ICA.

ICA is a statistical signal processing technique whose goal is to express a set of random variables as linear combinations of statistically independent variables. We observe k scalar variables $(d_1, d_2, \ldots, d_k)^{\mathrm{T}} \equiv \mathbf{d}$, which are assumed to be linear combinations of n unknown statistically independent variables $(s_1, s_2, \ldots, s_n)^{\mathrm{T}}$. We can express this mixing of the sources $\mathbf{s}$ as

$$\mathbf{d} = \mathbf{A}\mathbf{s}, \tag{2.1}$$

where **A** is an unknown $k \times n$ mixing matrix. The problem of ICA is to estimate both the mixing matrix **A** and the sources **s** using only the observation of the mixtures d_i. Using the feature extraction properties of ICA, the columns of **A** represent features, and s_i represent the amplitude of each feature in the observed mixtures **d**. These are the features in which we are interested.

In order to perform ICA, we first make a linear transformation of the observed mixtures

$$\mathbf{c} = \mathbf{M}\mathbf{d}. \tag{2.2}$$

These linearly transformed variables would be the outputs of the neurons in a neural network implementation, and **M**, the unmixing matrix or matrix of features, would be the weights. Two recent methods for performing ICA (Bell & Sejnowski, 1995; Amari, Cichocki, & Yang, 1996) involve maximizing the entropy of a nonlinear function of the transformed mixtures, $\sigma(\mathbf{c})$, and minimizing the mutual information of $\sigma(\mathbf{c})$ with respect to the transformation matrix, **M**, so that the components of $\sigma(\mathbf{c})$ are independent. These methods are, by their definition, multineuron algorithms and therefore do not fit well into the framework of this study.

The search for independent components relies on the fact that a linear mixture of two nongaussian distributions will become more gaussian than either of them. Thus, by seeking projections $c = (\mathbf{d} \cdot \mathbf{m})$ that maximize deviations from gaussian distribution, we recover the original (independent) signals. This explains the connection of ICA to the framework of exploratory projection pursuit (Friedman & Tukey, 1974; Friedman, 1987). In particular it holds for the kurtosis projection index, since a linear mixture will be less kurtotic than its original components.

Kurtosis and skewness have also been used for ICA as approximations of the negative entropy (Jones & Sibson, 1987). It remains to be seen if the basic assumption used in ICA, that the signals are made up of independent sources, is valid. The fact that different ICA algorithms, such as kurtosis and skewness maximization, yield different receptive fields could be an indication that the assumption is not completely valid.

3 Synaptic Modification Rules

In this section we outline the derivation for the learning rules in this study, using either the method from Projection Pursuit or ICA. Neural activity is assumed to be a positive quantity, so for biological plausibility, 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). σ' 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 that produced by natural scenes. We also

demonstrate later that the rectification makes little difference on learning rules that depend on even moments.

We study the following measures:

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

$$S_1 = E[c^3]/E^{1.5}[c^2]. \tag{3.1}$$

A maximization of this measure via gradient ascent gives

$$\begin{aligned}\nabla S_1 &= \frac{1}{{\Theta_M}^{1.5}} E\left[c\left(c - E[c^3]/E[c^2]\right)\sigma'\mathbf{d}\right] \\ &= \frac{1}{{\Theta_M}^{1.5}} E\left[c\left(c - E[c^3]/\Theta_M\right)\sigma'\mathbf{d}\right],\end{aligned} \tag{3.2}$$

where Θ_m is defined as $E[c^2]$.

Skewness 2. Another skewness measure is given by

$$S_2 = E[c^3] - E^{1.5}[c^2]. \tag{3.3}$$

This measure requires a stabilization mechanism, because it is not invariant under constant multiples of the activity c. We stabilize the rule by requiring that the vector of weights, which is denoted by m, has a fixed norm, say, $\|\mathbf{m}\| = \mathbf{1}$. The gradient of this measure is

$$\nabla S_2 = 3E\left[c^2 - c\sqrt{E[c^2]}\right] = 3E\left[c\left(c - \sqrt{\Theta_M}\right)\sigma'\mathbf{d}\right], \tag{3.4}$$

subject to the constraint $\|\mathbf{m}\| = \mathbf{1}$.

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

$$K_1 = E[c^4]/E^2[c^2] - 3. \tag{3.5}$$

This measure has a gradient of the form

$$\begin{aligned}\nabla K_1 &= \frac{1}{{\Theta_M}^2} E\left[c\left(c^2 - E[c^4]/E[c^2]\right)\sigma'\mathbf{d}\right] \\ &= \frac{1}{{\Theta_M}^2} E\left[c\left(c^2 - E[c^4]/\Theta_M\right)\sigma'\mathbf{d}\right].\end{aligned} \tag{3.6}$$

Kurtosis 2. As before, there is a similar form that requires some stabilization:

$$K_2 = E[c^4] - 3E^2[c^2]. \tag{3.7}$$

This measure has a gradient of the form

$$\begin{aligned}\nabla K_2 &= 4E\left[\left(c^3 - 3cE[c^2]\right)\sigma'\mathbf{d}\right] \\ &= 4E\left[c(c^2 - 3\Theta_M)\sigma'\mathbf{d}\right], \quad \|\mathbf{m}\| = \mathbf{1}.\end{aligned} \tag{3.8}$$

In all the above, the maximization of the measure can be used as a goal for projection seeking. 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. The multiplicative forms of both kurtosis and skewness do not require an extra stabilization constraint.

Kurtosis 2 and ICA. It has been shown (Hyvarinen & Oja, 1996) that kurtosis, defined as

$$K_2 = E\left[\mathbf{c^4}\right] - 3E^2\left[\mathbf{c^2}\right],$$

can be used for ICA. This can be seen by using the property of this kurtosis measure, $K_2(x_1 + x_2) = K_2(x_1) + K_2(x_2)$ for independent variables and defining $\mathbf{z} = \mathbf{A}^{\mathrm{T}}\mathbf{m}$. We then get

$$K_2(\mathbf{m}\cdot\mathbf{d}) \equiv \mathbf{K_2}(\mathbf{m}^{\mathrm{T}}\mathbf{d}) = \mathbf{K_2}(\mathbf{m}^{\mathrm{T}}\mathbf{A}\mathbf{s}) = \mathbf{K_2}(\mathbf{z}^{\mathrm{T}}\mathbf{s}) = \sum_{\mathbf{j=1}}^{\mathbf{n}} \mathbf{z_j^4 K_2}(\mathbf{s_j}). \tag{3.9}$$

The extremal points of equation 3.9 with respect to $\mathbf{z}$ under the constraint $E\left[(\mathbf{m}\cdot\mathbf{d})^{\mathbf{2}}\right] = 1$ occur when one component z_j of $\mathbf{z}$ is ± 1 and all the rest are zero (Delfosse & Loubaton, 1995). In other words, finding the extremal points of kurtosis leads to projections where $\mathbf{m}\cdot\mathbf{d} \equiv \mathbf{m}^{\mathrm{T}}\mathbf{d} = \mathbf{z}^{\mathrm{T}}\mathbf{s}$ equals, up to a sign, a single component s_j of $\mathbf{s}$. Thus, finding the extrema of kurtosis of the projections enables the estimation of the independent components individually, rather than all at once, as is done by other ICA rules. A full ICA code could be developed by introducing a lateral inhibition network, for example, but we restrict ourselves to the single neuron case here for simplicity.

Maximizing K_2 under the constraint $E\left[(\mathbf{m}\cdot\mathbf{d})^{\mathbf{2}}\right] = 1$, and defining the covariance matrix of the inputs $\mathbf{C} = \mathbf{E}\left[\mathbf{d}\mathbf{d}^{\mathrm{T}}\right]$, yields the following learning rule:

$$\mathbf{m} = \frac{\mathbf{2}}{\lambda}\left(\mathbf{C^{-1}E}\left[\mathbf{d}(\mathbf{m}\cdot\mathbf{d})^{\mathbf{3}}\right] - \mathbf{3m}\right). \tag{3.10}$$

This equation leads to an iterative fixed-point algorithm, which converges very quickly and works for both single-cell and network implementations (Hyvarinen & Oja, 1996).

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

$$\text{QBCM} = \frac{1}{3}E[c^3] - \frac{1}{4}E^2[c^2]. \tag{3.11}$$

Maximizing this form using gradient ascent gives the learning rule:

$$\nabla\text{QBCM} = E\left[c^2 - cE[c^2]\right] = E[c(c - \Theta_M)\sigma'\mathbf{d}]. \tag{3.12}$$

Unlike the measures S_2 and K_2 above, the QBCM 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).

4 Methods

We use 13×13 circular patches from 12 images of natural scenes as the visual environment. Two different types of preprocessing of the images are used 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 & Cooper, 1994). The second is a whitening filter, used to eliminate the second-order correlations (Oja, 1995; Bell & Sejnowski, 1995). Whitening the data in this way allows one to use learning rules that are dependent on higher moments of the data but are particularly sensitive to the second moment.

At each iteration of the learning, a 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 (i.e., K_2 and S_2) we also normalize the weights at each iteration.

For Oja's fixed-point algorithm, the learning was done in batches of 1000 patterns over which the expectation values were performed. However, the covariance matrix was calculated over the entire set of input patterns.

5 Results

5.1 Receptive Fields. The resulting receptive fields (RFs) formed are shown in Figures 1 and 2 for both the DOGed and whitened images, respec-

tively. Each RF shown was achieved using different random initial conditions for the weights. Every learning rule developed oriented RFs, though some were more sensitive to the preprocessing than others. The additive versions of kurtosis and skewness, K_2 and S_2, respectively, developed significantly different RFs in the whitened environment compared with the DOGged environment. The RFs in the whitened environment had higher spatial frequency and sampled from more orientations than the RFs in the DOGed environment. This behavior, as well as the resemblance of the RFs in the DOGed environment to those obtained from PCA (Shouval & Liu, 1996), suggest that these measures have a strong dependence on the second moment.

The multiplicative versions of kurtosis and skewness, K_1 and S_1 respectively, as well as QBCM, sampled from many orientations regardless of the preprocessing. The multiplicative skewness rule, S_1, gives RFs with lower spatial frequencies than either QBCM or the multiplicative kurtosis rule. This also disappears with the whitened inputs, which implies that the spatial frequency of the RFs is related to the strength of the dependence of the learning rule on the second moment. Example RFs using Oja's fixed-point ICA algorithm are also shown in Figure 2; not surprisingly, they look qualitatively similar to those found using the stochastic maximization of additive kurtosis, K_2.

The log of the output distributions for all of the rules has the double linear form, which implies a double exponential distribution. This distribution is one that we would consider sparse, but it would be difficult to compare the sparseness of the distributions merely on the appearance of the output distribution alone. In order to determine the sparseness of the code, we introduce a method for measuring it directly.

5.2 Structure Removal: Sensitivity to Outliers. Learning rules that are dependent on large polynomial moments, such as QBCM 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. Oversensitivity to outliers is considered to be undesirable in the statistical literature. However, in the case of a sparse code, the outliers, or the rare and interesting events, are what is important. The degree to which the neurons form a sparse code determines how much of the input distribution is required for maintaining the RF. This can be done in a straightforward and systematic fashion.

The procedure involves simply eliminating from the environment those patterns for which the neuron responds strongly. An example RF and some of the patterns that give that neuron strong responses are shown in Figure 3. These patterns tend to be the high-contrast edges and are thus the structure found in the image. The percentage of patterns that needs to be removed in order to cause a change in the RF gives a direct measure of the sparsity of the coding. The process of training a neuron, eliminating patterns that yield

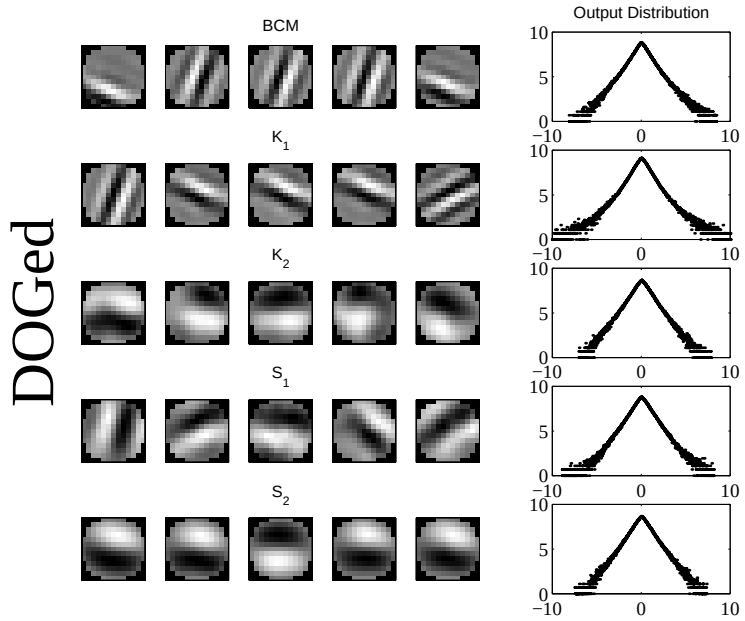

Figure 1: Receptive fields using DOGed image input obtained from learning rules maximizing (from top to bottom) the QBCM objective function, kurtosis (multiplicative), kurtosis (additive), skewness (multiplicative), and skewness (additive). Shown are five examples (left to right) from each learning rule, as well as the log of the normalized output distribution, before the application of the rectifying sigmoid.

high response, and retraining can be done recursively to remove structure sequentially from the input environment and to pick out the most salient features in the environment. The results are shown in Figure 4.

For QBCM and kurtosis, one need only eliminate less than one half of a percent of the input patterns to change the RF significantly. The changes that one can observe are orientation, phase, and spatial frequency changes. This is a very small percentage of the environment, which suggests that the neuron is coding the information in a very sparse manner. For the skewness maximization rule, more than 5% are needed to alter the RF properties, which implies a far less sparse coding.

To make this more precise, we introduce a normalized difference measure

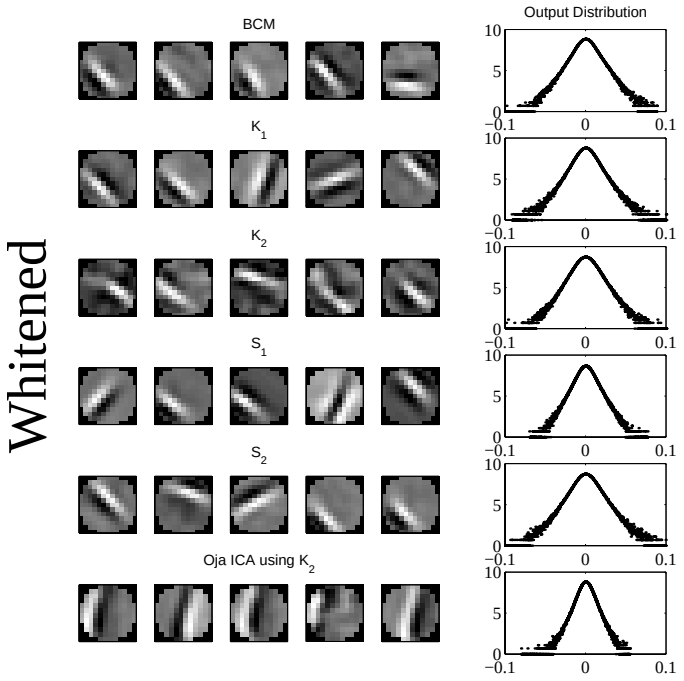

Figure 2: Receptive fields using whitened image input, obtained from learning rules maximizing (from top to bottom) the QBCM objective function, kurtosis (multiplicative), kurtosis (additive), skew (multiplicative), skewness (additive), and Oja's ICA rule based on the additive kurtosis measure. Shown are five examples (left to right) from each learning rule, as well as the log of the normalized output distribution, before the application of the rectifying sigmoid.

between two different RFs. If we take two weight vectors, $\mathbf{m_1}$ and $\mathbf{m_2}$, then the normalized difference between them is defined as

$$\mathcal{D} \equiv \frac{1}{4}\left(\frac{\mathbf{m_1}-\bar{\mathbf{m}}_\mathbf{1}}{\|\mathbf{m_1}\|}-\frac{\mathbf{m_2}-\bar{\mathbf{m}}_\mathbf{2}}{\|\mathbf{m_2}\|}\right)^2 \tag{5.1}$$

$$= \frac{1}{2}(1-\cos\alpha), \tag{5.2}$$

where α is the angle between the two vectors and $\bar{m}_i$ is the mean of the elements of the vector i. This measure is not sensitive to scale differences,

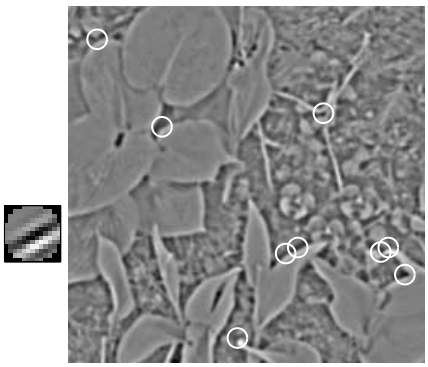

Figure 3: Patterns that yield high responses of a model neuron. The example receptive field is shown on the left. Some of the patterns that yield the strongest one-half percent of responses are labeled on the image on the right. These patterns are primarily the high-contrast edges.

because the vectors are divided by their norm, and it is not sensitive to scalar offset differences, because the mean of the vectors is subtracted. The measure has a value of zero for identical vectors, and a maximum value of one for orthogonal vectors.

Shown in Figure 5 is the normalized difference as a function of the percentage eliminated, for the different learning rules. Differences can be seen with as little as a tenth of a percent, but only changes of around a half-percent or above are visible as significant orientation, phase, or spatial frequency changes. Although both skewness and QBCM depend primarily on the third moment, QBCM behaves more like kurtosis with regard to projections from natural images.

Similar changes occur for both the BCM and kurtosis learning rules, and most likely occur under other rules that seek kurtotic projections. It is important to note, however, that patterns must be eliminated 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 would not 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 or possibly different position. Once enough input patterns are removed, the RF

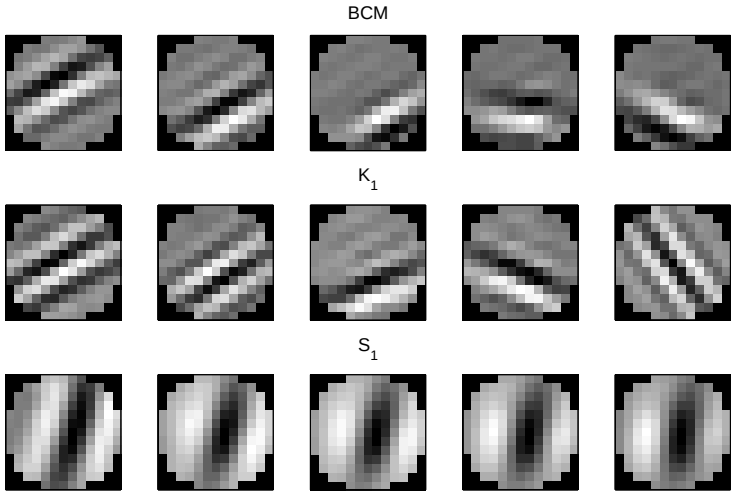

Figure 4: Receptive fields resulting from structure removal using the QBCM rule, the rule maximizing the multiplicative form of kurtosis and skewness. The RF on the far left for each rule was obtained in the normal input environment. The next RF to the right was obtained in a reduced input environment, 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 5% of the input patterns.

becomes oriented in a different direction. If the process is continued, all of the orientations and spatial locations would be obtained.

An objection may be made that the RFs formed are caused almost entirely by the application of the rectifying sigmoid. For odd-powered learning rules, the sigmoid is necessary to obtain oriented RFs because the distributions are approximately symmetric. This sigmoid is not needed for rules dependent on only the even-powered moments, such as kurtosis. Figure 6 demonstrates that the removal of the sigmoid and the removal of the mean from the moments calculations do not substantially affect the resulting RFs of the kurtosis rules.

The choice of 13×13 pixel RFs was biologically motivated and computationally less demanding than larger RFs formation. Figure 7 shows some 21×21 pixel RFs; it is clear that little difference is observed. One may notice that the length of the RF is larger than the lengths of the RFs found using

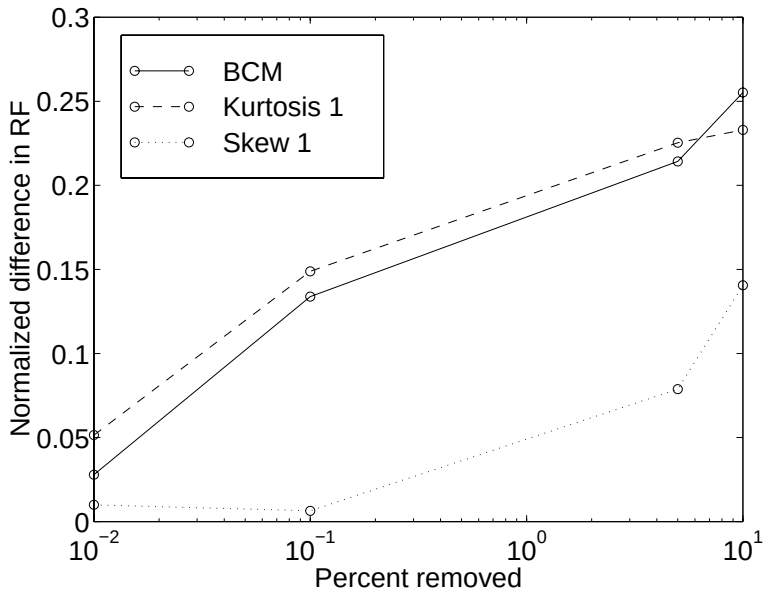

Figure 5: Normalized difference between RFs as a function of the percentage deleted in structure removal. The RFs were normalized, and mean zero, in order to neglect magnitude and additive constant changes. The maximum possible value of the difference is 1.

Olshausen and Field's or Bell and Sejnowski's algorithms. It is likely that events causing such elongated RFs are very rare, and thus lead to higher kurtosis. When additional constraints are imposed, such as finding a complete code, one ends up with less specialized RFs and, thus, less elongated RFs.

6 Discussion

This study compares several learning rules that have some statistical or biological motivation, or both. (For a related study discussing projection pursuit and BCM, see Press and Lee, 1997.) 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 because there are suggestions that a large family of linear filters can find kurtotic distributions (Ruderman, 1994).

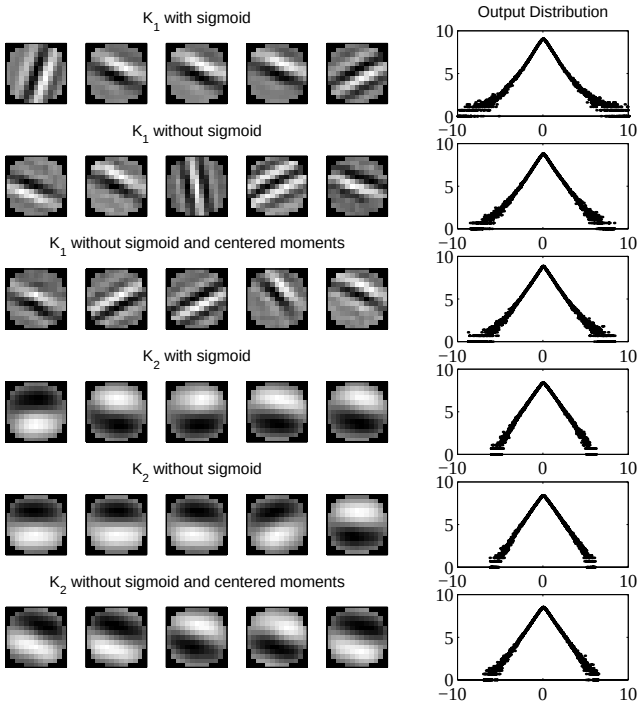

Figure 6: Receptive fields using DOGed image input, obtained from learning rules maximizing (from top to bottom) multiplicative form kurtosis with rectified outputs, nonrectified outputs, and nonrectified outputs with centered moments respectively, and additive form kurtosis with rectified outputs, nonrectified outputs, and nonrectified outputs with centered moments, respectively. Shown are five examples (left to right) from each learning rule and the corresponding output distribution.

- The single-cell ICA rule we considered used the subtractive form of kurtosis as a measure for deviation from gaussian distributions and achieved RFs qualitatively similar to other rules discussed.

- The QBCM and the multiplicative version of kurtosis are less sensitive to the second moments of the distribution and produce oriented RFs even when the data are not whitened. This is clear from the results

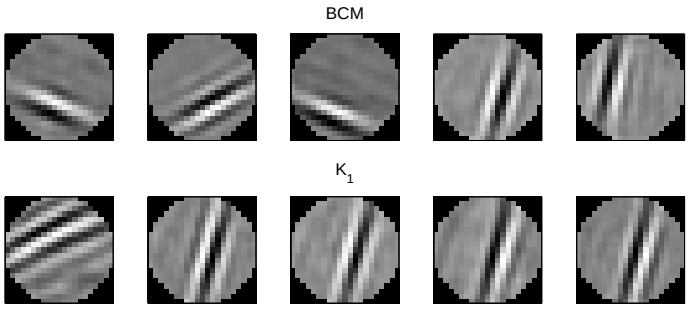

Figure 7: RFs using DOGed image input, obtained from the QBCM learning rule and the rule maximizing the multiplicative form of kurtosis.

about DOG-processed versus whitened inputs. The reduced sensitivity follows from the built-in second-order normalization that these rules have, kurtosis via division and BCM via subtraction. The subtractive version of kurtosis is sensitive and produces oriented RF only after sphering the data (Friedman, 1987; Field, 1994).

- Both QBCM and kurtosis are sensitive to the tails of the distribution. In fact,the RF changes due to elimination of the upper 1/2% portion of the distribution (see 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 of the kind of inputs the cell is most selective to (values below its highest 99% selectivity); these are inputs with same orientation but 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 studied further as it may reflect some fundamental instability of the kurtotic approaches.

- The skewness rules can also find oriented RFs. Their sensitivity to the upper parts of the distribution is not so dramatic, and thus the RFs do not change much when a few percent of the upper distribution are removed.

- Kurtotic rules can find high kurtosis in either symmetric or rectified distributions. This is not the case for QBCM rule, which requires rectified distributions.

- QBCM learning rule, which has been advocated as a projection index for finding multimodality in high-dimensional distribution, can find projections emphasizing high kurtosis when no cluster structure is present in the data. We have preliminary indications that the converse is not true: kurtosis measure does not perform well under distributions that are bi- or multimodal. This will be shown elsewhere.

Acknowledgments

This work was supported by the Office of Naval Research, the DANA Foundation, and the National Science Foundation.

References

Amari, S., Cichocki, A., & Yang, H. H. (1996). A new learning algorithm for blind signal separation. In G. Tesauro, D. Touretzky, & T. Leen (Eds.), *Advances in neural information processing systems, 8* (pp. 757–763).

Baddeley, R. J. (1996). An efficient code in V1? *Nature (News and Views), 381*, 560–561.

Bell, A. J., & Sejnowski, T. J. (1995). An information-maximisation approach to blind separation and blind deconvolution. *Neural Computation, 7*(6), 1129–1159.

Bell, A. J., & Sejnowski, T. J. (1997). The independent components of natural scenes are edge filters. *Vision Research, 37*(23), 3327–3338.

Bienenstock, E. L., Cooper, L. N., & Munro, P. W. (1982). Theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex. *Journal of Neuroscience, 2*, 32–48.

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

Daugman, J. G. (1988). Complete discrete 2D Gabor transforms by neural networks for image analysis and compression. *IEEE Transactions on ASSP, 36*, 1169–1179.

Delfosse, N., & Loubaton, P. (1995). Adaptive blind separation of independent sources: A deflation approach. *Signal Processing, 45*, 59–83.

Diaconis, P., & 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.

Friedman, J. H., & Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers, C(23)*, 881–889.

Huber, P. J. (1985). Projection pursuit (with discussion). *Annals of Statistics, 13*, 435–475.

Hyvarinen, A., & Oja, E. (1996). A fast fixed-point algorithm for independent component analysis. *Int. Journal of Neural Systems, 7*(6), 671–687.

Hyvarinen, A., & Oja, E. (1997). A fast fixed-point algorithm for independent component analysis. *Neural Computation, 9*(7), 671–687.

Intrator, N. (1990). A neural network for feature extraction. In D. S. Touretzky & R. P. Lippmann (Eds.), *Advances in neural information processing systems, 2* (pp. 719–726). San Mateo, CA: Morgan Kaufmann.

Intrator, N. (1996). Neuronal goals: Efficient coding and coincidence detection. In S. Amari, L. Xu, L. W. Chan, I. King, & K. S. Leung (Eds.), *Proceedings of ICONIP Hong Kong: Progress in Neural Information Processing* (Vol. 1, pp. 29–34). Berlin: Springer-Verlag.

Intrator, N. & Cooper, L. N. (1992). Objective function formulation of the BCM theory of visual cortical plasticity: Statistical connections, stability conditions. *Neural Networks, 5*, 3–17.

Jones, M. C., & Sibson, R. (1987). What is projection pursuit? (with discussion). *J. Roy. Statist. Soc., Ser. A, 150*, 1–36.

Kendall, M., & Stuart, A. (1977). *The advanced theory of statistics.* New York: Macmillan.

Law, C., & 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., & 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* (Tech. Rep. No. A26). Helsinki: Helsinki University, CS and Inf. Sci. Lab.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple cell receptvie field properties by learning a sparse code for natural images. *Nature, 381*, 607–609.

Press, W., & Lee, C. W. (1997). Searching for optimal value codes: Projection pursuit analysis of the statistical structure in natural scenes. In *The neurobiology of computation: Proceedings of the Fifth Annual Computation and Neural Systems Conference* (pp. 771–777). New York: Plenum.

Ruderman, D. L. (1994). The statistics of natural images. *Network, 5*(4), 517–548.

Shouval, H., & Liu, Y. (1996). Principal component neurons in a realistic visual environment. *Network, 7*(3), 501–515.

Zetzsche, C. (1997). Intrinsic dimensionality: How biological mechanisms exploit the statistical structure of natural images (talk given at the Natural Scene Statistics Meeting, Hancock, MA).

Received July 22, 1997; accepted March 2, 1998.

This article has been cited by:

1. Claudia Clopath, Lars Büsing, Eleni Vasilaki, Wulfram Gerstner. 2010. Connectivity reflects coding: a model of voltage-based STDP with homeostasis. *Nature Neuroscience* **13**:3, 344-352. [CrossRef]
2. Brian S. Blais, Harel Z. Shouval. 2009. Effect of correlated lateral geniculate nucleus firing rates on predictions for monocular eye closure versus monocular retinal inactivation. *Physical Review E* **80**:6. . [CrossRef]
3. Jochen Triesch. 2007. Synergies Between Intrinsic and Synaptic Plasticity MechanismsSynergies Between Intrinsic and Synaptic Plasticity Mechanisms. *Neural Computation* **19**:4, 885-909. [Abstract] [PDF] [PDF Plus]
4. Zhe Chen. 2005. Stochastic correlative firing for figure-ground segregation. *Biological Cybernetics* **92**:3, 192-198. [CrossRef]
5. S. Haykin, Z. Chen, S. Becker. 2004. Stochastic Correlative Learning Algorithms. *IEEE Transactions on Signal Processing* **52**:8, 2200-2209. [CrossRef]
6. Yuko Munakata, Jason Pfaffly. 2004. Hebbian learning and development. *Developmental Science* **7**:2, 141-148. [CrossRef]
7. A. Bazzani , D. Remondini , N. Intrator , G. C. Castellani . 2003. The Effect of Noise on a Class of Energy-Based Learning RulesThe Effect of Noise on a Class of Energy-Based Learning Rules. *Neural Computation* **15**:7, 1621-1640. [Abstract] [PDF] [PDF Plus]
8. Qian Yang, Xianglin Qi, Yunjiu Wang. 2001. A neural network model on self-organizing emergence of simple-cell receptive field with orientation selectivity in visual cortex. *Science in China Series C: Life Sciences* **44**:5, 469-478. [CrossRef]
9. Chris J. S. Webber . 2001. Predictions of the Spontaneous Symmetry-Breaking Theory for Visual Code Completeness and Spatial Scaling in Single-Cell Learning RulesPredictions of the Spontaneous Symmetry-Breaking Theory for Visual Code Completeness and Spatial Scaling in Single-Cell Learning Rules. *Neural Computation* **13**:5, 1023-1043. [Abstract] [PDF] [PDF Plus]
10. Chris J. S. Webber . 2000. Self-Organization of Symmetry Networks: Transformation Invariance from the Spontaneous Symmetry-Breaking MechanismSelf-Organization of Symmetry Networks: Transformation Invariance from the Spontaneous Symmetry-Breaking Mechanism. *Neural Computation* **12**:3, 565-596. [Abstract] [PDF] [PDF Plus]
11. Pamela Reinagel, Anthony Zador. 1999. Natural scene statistics at the centre of gaze. *Network: Computation in Neural Systems* **10**:4, 341-350. [CrossRef]
12. G Castellani, N Intrator, H Shouval, L Cooper. 1999. Solutions of the BCM learning rule in a network of lateral interacting nonlinear neurons. *Network: Computation in Neural Systems* **10**:2, 111-121. [CrossRef]
13. Harel Z ShouvalHebb Synapses: Modeling of Neuronal Selectivity . [CrossRef]