Communicated by Geoffrey Hinton

Feature Extraction Using an Unsupervised Neural Network

Nathan Intrator
Center for Neural Science, Brown University
Providence, RI 02912 USA

A novel unsupervised neural network for dimensionality reduction that seeks directions emphasizing multimodality is presented, and its connection to exploratory projection pursuit methods is discussed. This leads to a new statistical insight into the synaptic modification equations governing learning in Bienenstock, Cooper, and Munro (BCM) neurons (1982). The importance of a dimensionality reduction principle based solely on distinguishing features is demonstrated using a phoneme recognition experiment. The extracted features are compared with features extracted using a backpropagation network.

1 Introduction

When a classification of high-dimensional vectors is sought, the *curse of dimensionality* (Bellman 1961) becomes the main factor affecting the classification performance. The curse of dimensionality is due to the inherent sparsity of high-dimensional spaces, implying that, in the absence of simplifying assumptions, the amount of training data needed to get reasonably low variance estimators is ridiculously high. This has led many researchers in recent years to construct methods that specifically avoid this problem (see Geman *et al.* 1991 for review in the context of neural networks). One approach is to assume that important structure in the data actually lies in a much smaller dimensional space, and therefore try to reduce the dimensionality before attempting the classification. This approach can be successful if the dimensionality reduction/feature extraction method loses as little relevant information as possible in the transformation from the high-dimensional space to the low-dimensional one.

Performing supervised feature extraction using the class labels is sensitive to the dimensionality in a similar manner to a high-dimensional classifier, and may result in a strong bias to the training data leading to poor generalization properties of the resulting classifier (Barron and Barron 1988).

A general class of unsupervised dimensionality reduction methods, called exploratory projection pursuit, is based on seeking *interesting* projections of high-dimensional data points (Kruskal 1972; Friedman and

Tukey 1974; Friedman 1987; Huber 1985, for review). The notion of interesting projections is motivated by an observation made by Diaconis and Freedman (1984) that for most high-dimensional clouds, most low-dimensional projections are approximately normal. This finding suggests that important information in the data is conveyed in those directions whose single-dimensional projected distribution is far from gaussian. Various projection indices differ on the assumptions about the nature of deviation from normality, and in their computational efficiency. Friedman (1987) argues that the most computationally efficient measures are based on polynomial moments. However, polynomial moments heavily emphasize departure from normality in the tails of the distribution (Huber 1985). Moreover, although many synaptic plasticity models are based on second-order statistics and lead to extraction of the principal components (Sejnowski 1977; von der Malsburg 1973; Oja 1982; Miller 1988; Linsker 1988), second-order polynomials are not sufficient to characterize the important features of a distribution (see examples in Duda and Hart 1973, p. 212). This suggests that in order to use polynomials for measuring deviation from normality, higher order polynomials are required, and care should be taken to avoid their oversensitivity to outliers. In this paper, the observation that high-dimensional clusters translate to multimodal low-dimensional projections is used for defining a measure of multimodality for seeking interesting projections. In some special cases, where the data are known in advance to be bimodal, it is relatively straightforward to define a good projection index (Hinton and Nowlan 1990), however, when the structure is not known in advance, defining a general multimodal measure of the projected data is not straightforward, and will be discussed in this paper.

There are cases in which it is desirable to make the projection index invariant under certain transformations, and maybe even remove second-order structure (see Huber 1985 for desirable invariant properties of projection indices). In those cases it is possible to make such transformations beforehand (Friedman 1987), and then assume that the data possess these invariant properties.

2 Feature Extraction Using ANN

In this section, the intuitive idea presented above is used to form a statistically plausible objective function whose minimization will find those projections having a single-dimensional projected distribution that is far from gaussian. This is done using a loss function that has an expected value that leads to the desired projection index. Mathematical details are given in Intrator (1990).

Before presenting our version of the loss function, we review some necessary notation and assumptions. Consider a neuron with input vector $x = (x_1, \ldots, x_N)$, synaptic weight vector $m = (m_1, \ldots, m_N)$, both in

R^N, and activity (in the linear region) $c = x \cdot m$. Define the threshold $\Theta_m = E[(x \cdot m)^2]$, and the functions $\hat{\phi}(c, \Theta_m) = c^2 - (2/3)c\Theta_m$, $\phi(c, \Theta_m) = c^2 - (4/3)c\Theta_m$. The ϕ function has been suggested as a biologically plausible synaptic modification function to explain visual cortical plasticity (Bienenstock *et al.* 1982). Θ_m is a dynamic threshold that will be shown later to have an effect on the sign of the synaptic modification. The input x, which is a stochastic process, is assumed to be of Type II φ mixing,[1] bounded, and piecewise constant. These assumptions are plausible, since they represent the closest continuous approximation to the usual training algorithms, in which training patterns are presented at random. The φ mixing property allows for some time dependency in the presentation of the training patterns. These assumptions are needed for the approximation of the resulting deterministic gradient descent by a stochastic one (Intrator and Cooper 1991). For this reason we use a *learning rate* μ that has to decay in time so that this approximation is valid.

We want to base the projection index on polynomial moments of low order, and to use the fact that a projection that leads to a bimodal distribution is already interesting, and any additional mode in the projected distribution should make the projection even more interesting. With this in mind, consider the following family of loss functions that depends on the synaptic weight vector m and on the input x;

$$L_m(x) = -\mu \int_0^{(x \cdot m)} \hat{\phi}(s, \Theta_m)\, ds = -\frac{\mu}{3}\{(x \cdot m)^3 - E[(x \cdot m)^2](x \cdot m)^2\}$$

The motivation for this loss function can be seen in Figure 1, which represents the $\hat{\phi}$ function and the associated loss function $L_m(c)$. For simplicity the loss for a fixed threshold Θ_m and synaptic vector m can be written as $L_m(c) = -(\mu/3)c^2(c - \Theta_m)$, where $c = (x \cdot m)$.

The graph of the loss function shows that for any fixed m and Θ_m, the loss is small for a given input x, when either $c = x \cdot m$ is close to zero, or when $x \cdot m$ is larger than Θ_m. Moreover, the loss function remains negative for $(x \cdot m) > \Theta_m$, therefore any kind of distribution at the right-hand side of Θ_m is possible, and the preferred ones are those that are concentrated further from Θ_m.

It remains to be shown why it is not possible that a minimizer of the average loss will be such that all the mass of the distribution will be concentrated on one side of Θ_m. This can not happen because the threshold Θ_m is dynamic and depends on the projections in a nonlinear way, namely, $\Theta_m = E(x \cdot m)^2$. This implies that Θ_m will always move itself to a position such that the distribution will never be concentrated at only one of its sides.

The risk (expected value of the loss) is given by

$$R_m = -\frac{\mu}{3}\{E[(x \cdot m)^3] - E^2[(x \cdot m)^2]\}$$

[1]The φ mixing property specifies the dependency of the future of the process on its past.

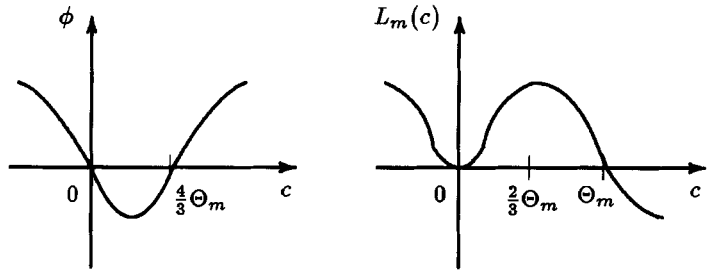

Figure 1: The function ϕ and the loss functions for a fixed m and Θ_m.

Since the risk is continuously differentiable, its minimization can be achieved via a gradient descent method with respect to m, namely

$$\frac{dm_i}{dt} = -\frac{\partial}{\partial m_i} R_m = \mu \, E[\phi(x \cdot m, \Theta_m) x_i]$$

The resulting differential equations give a modified version of the law governing synaptic weight modification in the BCM theory for learning and memory (Bienenstock *et al.* 1982). This theory was presented to account for various experimental results in visual cortical plasticity. The modification lies in the way the threshold Θ_m is calculated. In the original form this threshold was $\Theta_m = E^p(c)$ for $p > 1$, while in the current form $\Theta_m = E(c^p)$ for $p > 1$. The latter takes into account the variance of the activity (for $p = 2$) and therefore is always positive; this ensures stability even when the average of the inputs is zero. The biological relevance of the theory has been extensively studied (Bear *et al.* 1987; Bear and Cooper 1988) and it was shown that the theory is in agreement with the classical deprivation experiments (Clothiaux *et al.* 1991).

The fact that the distribution has part of its mass on both sides of Θ_m makes this loss a plausible projection index that seeks multimodalities. However, we still need to reduce the sensitivity of the projection index to outliers, and for full generality, allow any projected distribution to be shifted so that the part of the distribution that satisfies $c < \Theta_m$ will have its mode at zero. The oversensitivity to outliers is addressed by considering a nonlinear neuron in which the neuron's activity is defined to be $c = \sigma(x \cdot m)$, where σ usually represents a smooth sigmoidal function. A more general definition that would allow symmetry breaking of the projected distributions, as well as provide a solution to the second problem raised above, and will still be consistent with the statistical formulation, is $c = \sigma(x \cdot m - \alpha)$, for an arbitrary threshold α. The threshold α can be found by using gradient descent as well. For the nonlinear neuron, Θ_m is defined

to be $\Theta_m = E[\sigma^2(x \cdot m)]$. The loss function is given by

$$L_m(x) = -\mu \int_0^{\sigma(x \cdot m)} \hat{\phi}(s, \Theta_m) ds = -\frac{\mu}{3} \{\sigma^3(x \cdot m) - E[\sigma^2(x \cdot m)]\sigma^2(x \cdot m)\}$$

The gradient of the risk becomes

$$-\nabla_m R_m = \mu \, E[\phi(\sigma(x \cdot m), \Theta_m) \, \sigma' x]$$

where σ' represents the derivative of σ at the point $(x \cdot m)$. Note that the multiplication by σ' reduces sensitivity to outliers of the differential equation since for outliers σ' is close to zero. The gradient descent is valid, provided that the risk is bounded from below.

Based on this formulation, a network of Q identical nodes may be constructed. All the neurons in this network receive the same input and inhibit each other, so as to extract several features in parallel. The relation between this network and the network studied by Cooper and Scofield (1988) is discussed in Intrator and Cooper (1991). The activity of neuron k in the network is defined as $c_k = \sigma(x \cdot m_k - \alpha_k)$, where m_k is the synaptic weight vector of neuron k, and α_k is its threshold. The *inhibited* activity and threshold of the kth neuron are given by $\tilde{c}_k = c_k - \eta \sum_{j \neq k} c_j$, $\tilde{\Theta}_m^k = E[\tilde{c}_k^2]$. A more general inhibitory pattern such as a Mexican hat is possible with minor changes in the mathematical details.

We omit the derivation of the synaptic modification equations, and present only the resulting stochastic modification equations for a synaptic vector m_k in a lateral inhibition network of nonlinear neurons:

$$\dot{m}_k = \mu \, [\phi(\tilde{c}_k, \tilde{\Theta}_m^k)\sigma'(\tilde{c}_k) - \eta \sum_{j \neq k} \phi(\tilde{c}_j, \tilde{\Theta}_m^j)\sigma'(\tilde{c}_j)]x$$

The lateral inhibition network performs a direct search of Q-dimensional projections in parallel, and therefore may find a richer structure that a step wise approach may miss (see example 14.1 in Huber 1985).

3 Comparison with Other Feature Extraction Methods

The above feature extraction method has been applied so far to various high-dimensional classification problems: extracting rotation invariant features from 3D wire-like objects (Intrator and Gold 1991) based on a set of sophisticated psychophysical experiments (Edelman and Bülthoff 1991); feature extraction from the TIMIT speech data base using Lyon's Cochlea model (Intrator and Tajchman 1991). The dimensionality of the feature extraction problem for these experiments was 3969 and 5500 dimensions, respectively. It is surprising that a very moderate amount of training data was needed for extracting robust features as will be shown below. In this section we briefly describe a linguistically motivated feature extraction experiment from stop consonants. We compare classification performance of the proposed method to a network that performs

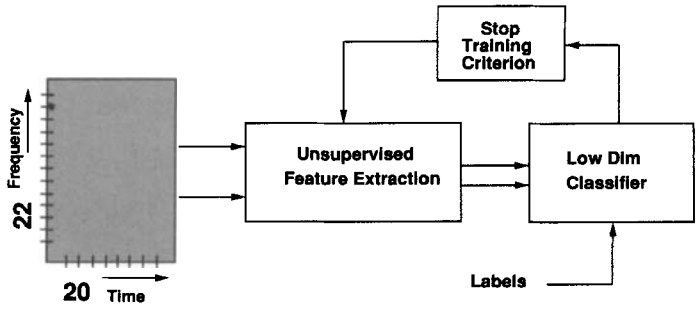

Figure 2: Low-dimensional classifier is trained on features extracted from the high-dimensional data. Training of the feature extraction network stops when the misclassification rate drops below a predetermined threshold on either the same training data (cross-validatory test) or on different testing data.

dimensionality reduction based on minimization of misclassification error (using backpropagation with MSE criterion). In the latter we regard the hidden unit representation as a new reduced feature representation of the input space. Classification on the new feature space was done using backpropagation.[2]

The unsupervised feature extraction/classification method is presented in Figure 2. The pixel images corresponding to speech data, are shown in Figure 3. Similar approaches using the RCE and backpropagation network have been carried out by Reilly *et al.* (1988).

The following describes the linguistic motivation of the experiment. Consider the six stop consonants [p,k,t,b,g,d], which have been a subject of recent research in evaluating neural networks for phoneme recognition (see review in Lippmann 1989). According to phonetic feature theory, these stops possess several common features, but only two distinguishing phonetic features, place of articulation and voicing (see Lieberman and Blumstein 1988, for a review and related references on phonetic feature theory). This theory suggests an experiment in which features extracted from unvoiced stops can be used to distinguish place of articulation in voiced stops as well. It is of interest if these features can be found from a single speaker, how sensitive they are to voicing and whether they are speaker invariant.

The speech data consists of 20 consecutive time windows of 32 msec with 30 msec overlap, aligned to the beginning of the burst. In each time window, a set of 22 energy levels is computed. These energy levels cor-

[2]See Intrator (1990) for comparison with principal components feature extraction and with k-NN as a classifier.

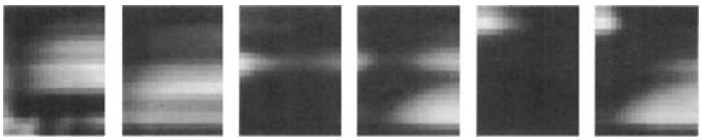

Figure 3: An average of the six stop consonants followed by the vowel [a]. Their order from left to right [pa] [ba] [ka] [ga] [ta] [da]. Time increases from the burst release on the X axis, and frequency increases on the Y axis. Brighter areas correspond to stronger energy.

respond to Zwicker critical band filters (Zwicker 1961). The consonant–vowel (CV) pairs were pronounced in isolation by native American speakers (two male BSS and LTN, and one female JES.) Additional details on biological motivation for the preprocessing, and linguistic motivation related to child language acquisition can be found in Seebach (1990). An average (over 25 tokens) of the six stop consonants followed by the vowel [a] is presented in Figure 3. All the images are smoothed using a moving average. One can see some similarities between the voiced and unvoiced stops especially in the upper left corner of the image (high frequencies beginning of the burst) and the radical difference between them in the low frequencies.

In the experiments reported here, five features were extracted from the 440 dimension original space. Although the dimensionality reduction methods were trained only with the unvoiced tokens of a single speaker, the classifier was trained on (five-dimensional) voiced and unvoiced data from the other speakers as well.

The classification results, which are summarized in Table 1, show that the backpropagation network does well in finding structure useful for classification of the trained data, but this structure is more sensitive to voicing. Classification results using a BCM network suggest that for this specific task structure that is less sensitive to voicing can be extracted, even though voicing has significant effects on the speech signal itself. The results also suggest that these features are more speaker invariant.

The difference in performance between the two feature extractors may be partially explained by looking at the synaptic weight vectors (images) extracted by both methods (Fig. 4): For the backpropagation feature extraction it can be seen that although five units were used, less features were extracted. One of the main distinctions between the unvoiced stops in the training set is the high frequency burst at the beginning of the consonant (the upper left corner). The backpropagation method concentrated mainly on this feature, probably because it is sufficient to base the recognition of the training set on this feature, and the fact that training

Table 1: Percentage of Correct Classification of Place of Articulation in Voiced and Unvoiced Stops.

	Place of articulation classification (B-P)	
	B-P (%)	BCM (%)
BSS /p,k,t/	100	100
BSS /b,g,d/	83.4	94.7
LTN /p,k,t/	95.6	97.7
LTN /b,g,d/	78.3	93.2
JES (both)	88.0	99.4

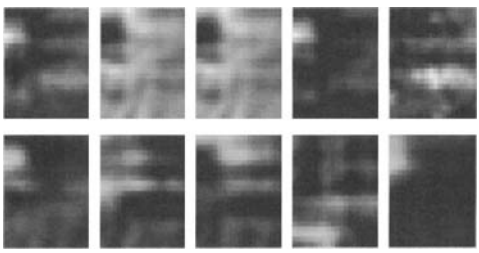

Figure 4: Synaptic weight images of the five hidden units of backpropagation (top), and the five BCM neurons (bottom).

stops when misclassification error falls to zero. On the other hand, the BCM method does not try to reduce the misclassification error and is able to find a richer, linguistically meaningful structure, containing burst locations and format tracking of the three different stops that allowed a better generalization to other speakers and to voiced stops.

The network and its training paradigm present a different approach to speaker independent speech recognition. In this approach the speaker variability problem is addressed by training a network that concentrates mainly on the distinguishing features of a single speaker, as opposed to training a network that concentrates on both the distinguishing and common features, on multispeaker data.

Acknowledgments

I wish to thank Leon N. Cooper for suggesting the problem and for providing many helpful hints and insights. Geoff Hinton made invaluable comments. The application of BCM to speech is discussed in more detail in Seebach (1991) and in a forthcoming article (Seebach and Intrator, in press). Charles Bachmann assisted in running the backpropagation experiments.

Research was supported by the National Science Foundation, the Army Research Office, and the Office of Naval Research.

References

Barron, A. R., and Barron, R. L. 1988. Statistical learning networks: A unifying view. In *Computing Science and Statistics: Proc. 20th Symp. Interface*, E. Wegman, ed., pp. 192–203. American Statistical Association, Washington, DC.

Bear, M. F., and Cooper, L. N 1988. Molecular mechanisms for synaptic modification in the visual cortex: Interaction between theory and experiment. In *Neuroscience and Connectionist Theory*, M. Gluck and D. Rumelhart, eds., pp. 65–94. Lawrence Erlbaum, Hillsdale, NJ.

Bear, M. F., Cooper, L. N, and Ebner, F. F. 1987. A physiological basis for a theory of synapse modification. *Science* **237**, 42–48.

Bellman, R. E. 1961. *Adaptive Control Processes.* Princeton University Press, Princeton, NJ.

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

Clothiaux, E. E., Cooper, L. N, and Bear, M. F. 1991. Synaptic plasticity in visual cortex: Comparison of theory with experiment. *J. Neurophysiol.* To appear.

Cooper, L. N, and Scofield, C. L. 1988. Mean-field theory of a neural network. *Proc. Natl. Acad. Sci. U.S.A.* **85**, 1973–1977.

Diaconis, P., and Freedman, D. 1984. Asymptotics of graphical projection pursuit. *Ann. Statist.* **12**, 793–815.

Duda, R. O., and Hart, P. E. 1973. *Pattern Classification and Scene Analysis.* John Wiley, New York.

Edelman, S., and Bülthoff, H. H. 1991. Canonical views and the representation of novel three-dimensional objects. To appear.

Friedman, J. H. 1987. Exploratory projection pursuit. *J. Amer. Statist. Assoc.* **82**, 249–266.

Friedman, J. H., and Tukey, J. W. 1974. A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Comput.* **C**(23), 881–889.

Geman, S., Bienenstock, E., and Doursat, R. 1991. Neural networks and the bias-variance dilemma. To appear.

Hinton, G. E., and Nowlan, S. J. 1990. The bootstrap Widrow-Hoff rule as a cluster-formation algorithm. *Neural Comp.* **2**(3), 355–362.

Huber, P. J. 1985. Projection pursuit (with discussion). *Ann. Statist.* **13**, 435–475.

Intrator, N., 1990. Feature extraction using an unsupervised neural network. In *Proceedings of the 1990 Connectionist Models Summer School*, D. S. Touretzky, J. L. Ellman, T. J. Sejnowski, and G. E. Hinton, eds., pp. 310–318. Morgan Kaufmann, San Mateo, CA.

Intrator, N., and Cooper, L. N 1991. Objective function formulation of the BCM theory of visual cortical plasticity: Statistical connections, stability conditions. *Neural Networks*. To appear.

Intrator, N., and Gold, J. I. 1991. Three-dimensional object recognition of gray level images: The usefulness of distinguishing features. To appear.

Intrator, N., and Tajchman, G. 1991. Supervised and unsupervised feature extraction from a cochlear model for speech recognition. In *Neural Networks for Signal Processing — Proceedings of the 1991 IEEE Workshop*, B. H. Juang, S. Y. Kung, and C. A. Kamm, eds., pp. 460–469.

Kruskal, J. B. 1972. Linear transformation of multivariate data to reveal clustering. In *Multidimensional Scaling: Theory and Application in the Behavioral Sciences, I, Theory*, R. N. Shepard, A. K. Romney, and S. B. Nerlove, eds., pp. 179–191. Seminar Press, New York and London.

Lieberman, P., and Blumstein, S. E. 1988. *Speech Physiology, Speech Perception, and Acoustic Phonetics*. Cambridge University Press, Cambridge.

Linsker, R. 1988. Self-organization in a perceptual network. *IEEE. Comp.* **88**, 105–117.

Lippmann, R. P. 1989. Review of neural networks for speech recognition. *Neural Comp.* **1**, 1–38.

Miller, K. D. 1988. Correlation-based models of neural development. In *Neuroscience and Connectionist Theory*, M. Gluck and D. Rumelhart, eds., pp. 267–353. Lawrence Erlbaum, Hillsdale, NJ.

Oja, E. 1982. A simplified neuron model as a principal component analyzer. *Math. Biol.* **15**, 267–273.

Reilly, D. L., Scofield, C. L., Cooper, L. N, and Elbaum, C. 1988. Gensep: A multiple neural network with modifiable network topology. *INNS Conf. Neural Networks*.

Seebach, B. S. 1991. Evidence for the development of phonetic property detectors in a neural net without innate knowledge of linguistic structure. Ph.D. dissertation, Brown University.

Seebach, B. S., and Intrator, N. A neural net model of perinatal inductive acquisition of phonetic features.

Sejnowski, T. J. 1977. Storing covariance with nonlinearly interacting neurons. *J. Math. Biol.* **4**, 303–321.

von der Malsburg, C. 1973. Self-organization of orientation sensitivity cells in the striate cortex. *Kybernetik* **14**, 85–100.

Zwicker, E. 1961. Subdivision of the audible frequency range into critical bands (frequenzgruppen). *J. Acoust. Soc. Am.* **33**(2): 248.

Received 18 March 1991; accepted 20 May 1991.