Communicated by Steven Nowlan

A Multiple Cause Mixture Model for Unsupervised Learning

Eric Saund
Xerox Palo Alto Research Center, 3333 Coyote Hill Rd., Palo Alto, CA 94304 USA

This paper presents a formulation for unsupervised learning of clusters reflecting multiple causal structure in binary data. Unlike the "hard" k-means clustering algorithm and the "soft" mixture model, each of which assumes that a single hidden event generates each data point, a multiple cause model accounts for observed data by combining assertions from many hidden causes, each of which can pertain to varying degree to any subset of the observable dimensions. We employ an objective function and iterative gradient descent learning algorithm resembling the conventional mixture model. A crucial issue is the *mixing function* for combining beliefs from different cluster centers in order to generate data predictions whose errors are minimized both during recognition and learning. The mixing function constitutes a prior assumption about underlying structural regularities of the data domain; we demonstrate a weakness inherent to the popular weighted sum followed by sigmoid squashing, and offer alternative forms of the nonlinearity for two types of data domain. Results are presented demonstrating the algorithm's ability successfully to discover coherent multiple causal representations in several experimental data sets.

1 Introduction

The objective of unsupervised learning is to identify patterns or features reflecting regularities in data. Algorithms vary in the assumptions they make about the underlying structural characteristics of the data domain, and they vary therefore in the nature of the patterns that can be discovered. This paper addresses unsupervised learning of *multiple cause* clusters in binary data, and identifies the *mixing function*, corresponding to a neural network's unit activation function, as an appropriate site at which to install prior domain knowledge of the ways in which hidden processes causally interact to generate observed data.

A *multiple-cause* model differs from a *single-cause* model in that it permits more than one hidden cluster-center to become fully "active" in accounting for an observed data point. The well-known k-means clustering algorithm, and its "softer" variant, the standard mixture model (Duda and Hart 1973; Nowlan 1990), are both single cause unsupervised

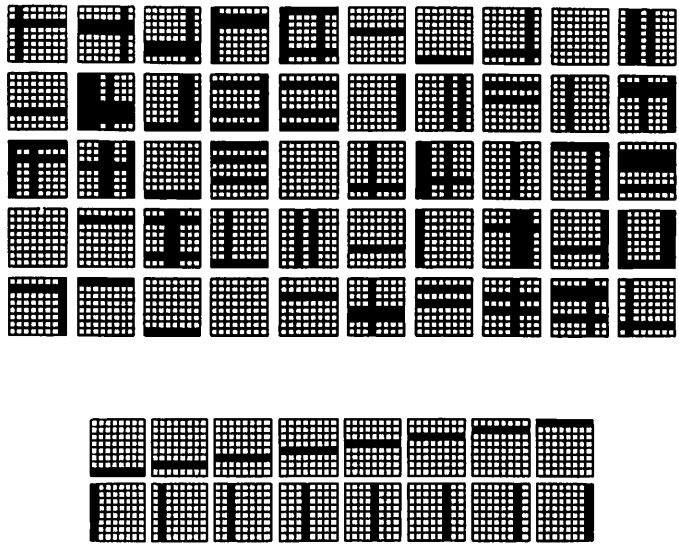


Figure 1: (a) Samples from a data set designed by Földiák (1990) consisting of horizontal and vertical lines in an 8×8 grid, each painted black with probability 1/8. (b) The ideal multiple cause representation for this data set consists of 16 independent components.

learning models by virtue of a winner-take-all step, or, alternatively a normalization step, such that cluster-center activities are constrained to sum to unity. Under a multiple cause model (also known as *componential* or *factorial* representation), cluster-centers are permitted to narrow their descriptive scope to only certain subspaces of the full data space, and therefore to share responsibility in accounting for observed data. The advantage of a multiple cause model is that a relatively small number of hidden variables can be applied combinatorially to generate a large data set. Figure 1 illustrates this with a test data set generated by the independent actions of 16 underlying components appearing as horizontal and vertical lines (Földiák 1990).

In the example of Figure 1, hidden causes corresponding to horizontal and vertical lines interact in a particularly simple way such that data pixels occurring at line intersections remain black. This mode of causal interaction—and other more complex modes of interaction found in other data sets—makes certain implications about the mixing functions appropriate for learning the patterns reflected by the underlying causal processes.

2 Common Architecture for Unsupervised Learning Models

A large class of single cause and multiple cause unsupervised learning models can be cast in the architecture shown in Figure 2. A binary vector $d_i \equiv (d_{i,1}, d_{i,2}, \ldots d_{i,j}, \ldots d_{i,J})$ is presented at the data layer, and a *measurement*, or response vector $m_i \equiv (m_{i,1}, m_{i,2}, \ldots m_{i,k}, \ldots m_{i,K})$ is computed at the encoding layer using "weights" $c_{j,k}$ associating activity at data dimension j with activity at hidden cluster-center k. Any activity pattern at the encoding layer can be turned around to compute a *prediction* vector $r_i \equiv (r_{i,1}, r_{i,2}, \ldots r_{i,j}, \ldots r_{i,J})$. Different models employ different functions for performing the measurement and prediction mappings, and give different interpretations to the weights. For example, under the k-means model the weights correspond to locations of cluster-centers in the data space; measurement is performed by computing distance from an observed data vector to each cluster-center, and then performing winner-take-all. Prediction of a data point is simply the vector C_k of the single active (kth) cluster-center. Likewise interpretations can be given to the mixture model, principal components methods (Bourlard and Kamp 1988; Sanger 1989) (including encoder networks trained by backpropagation, in which measurement weights may differ from prediction weights), and the Harmonium Boltzmann Machine (Freund and Haussler 1992). Common to all these models is a learning procedure that attempts to optimize an objective function on errors between data vectors in a training set, and predictions of these data vectors under their respective responses at the encoding layer.

Földiák (1990) and more recently Zemel (1993) has shown that under some circumstances appropriate multiple cause representations can be induced for data sets such as Figure 1 by incorporating auxiliary con-

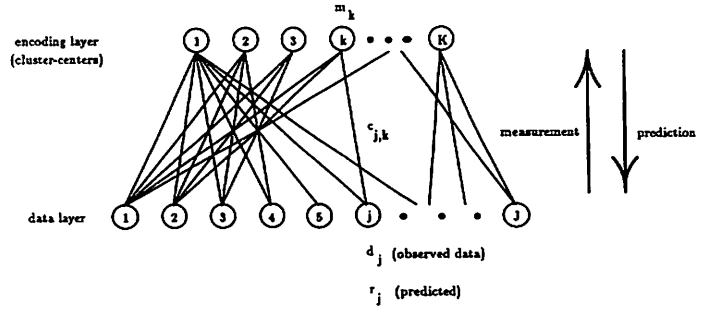


Figure 2: Architecture underlying a large class of unsupervised learning models.

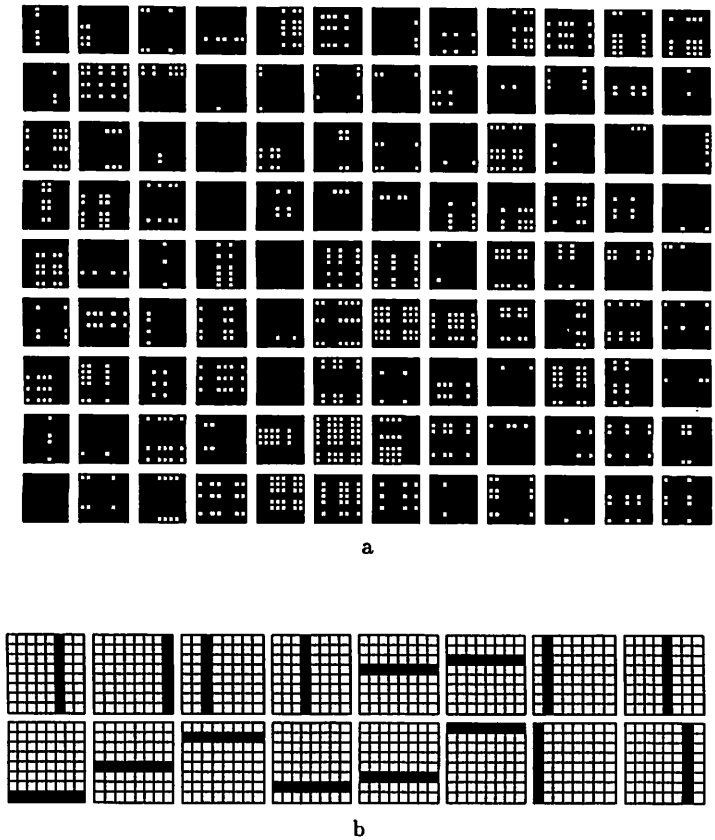


Figure 3: (a) Data set consisting of horizontal and vertical lines occurring with probability 0.625. (b) Multiple cause representation for 2000 randomly generated data points of this type discovered by the multiple cause mixture model using the soft OR mixing function.

straints on activity patterns at the encoding layer. In particular, in accordance with a 1/8 probability for each horizontal or vertical line to appear, they incorporate a *sparseness* assumption taking form as pressure for few hidden units to become active at one time. While sparseness is motivated by various theoretical considerations, it is an inappropriate assumption for the data of Figure 3, which shares the same underlying structure of independent horizontal and vertical lines, except here each line occurs with probability 0.625. We turn instead to the mixing function as a site

at which to achieve greater leverage in domain-dependent assumptions about the behavior of the underlying causes.

3 Imaging Models and Voting Rules

Mixing functions may be conceived metaphorically in two ways that are useful in designing them to reflect domain-specific modes of causal interaction. First, a mixing function is equivalent to an *imaging model* in the sense of digital typography and graphics (Warnock and Wyatt 1982); an imaging model specifies how layers of "color" combine on a surface to give rise to some resulting visible color. The imaging model corresponding to the horizontal and vertical lines data is known as a WRITE-BLACK imaging model. By default, prediction layer activities are OFF, corresponding to white pixels. Activity at a hidden unit colors ON, or black, into a row or column of pixels. Furthermore, pixels falling at the intersections of ON horizontal and vertical lines remain ON. The WRITE-BLACK imaging model therefore corresponds to a disjunctive—logical OR—mode of causal interaction.

A second way to view a mixing function is as a voting rule. Each hidden unit may be considered as holding some opinion or belief about the value of each prediction unit to which it is connected, arising from the hidden unit's degree of activity and its connection weight to the prediction unit. The purpose of a mixing function is, for each prediction unit, to collect the possibly conflicting beliefs and negotiate a net prediction output. Corresponding to the WRITE-BLACK imaging model is a disjunctive voting scheme in which hidden units are allowed to either abstain, or else vote ON, whereby any single hidden unit voting ON is sufficient to drive that prediction unit ON.

An appropriate mixing function for WRITE-BLACK type multiple cause binary data domains is therefore based on disjunctive voting by the unobserved causes. To learn the actual mappings between causes and data patterns, however, it is necessary to "soften" the logical OR voting rule so that learning may be achieved by performing gradient descent in weight space. This is accomplished by linearly interpolating the boolean OR function, which can be shown to yield the soft disjunctive mixing function given by the expression,

$$_{wb}r_{i,j} = 1 - \prod_k (1 - m_{i,k}c_{j,k}) \tag{3.1}$$

(see Fig. 4).

Using this mixing function, the 16 independent horizontal and vertical lines are discovered both for data of Figure 1 and of Figure 3 in which hidden causes are active on average in over half of the data samples. Figure 5 displays underlying image fragments discovered to decompose test data consisting of random spline curve images (Hinton and Zemel 1994). Qualitatively similar fragments are found whether one or several

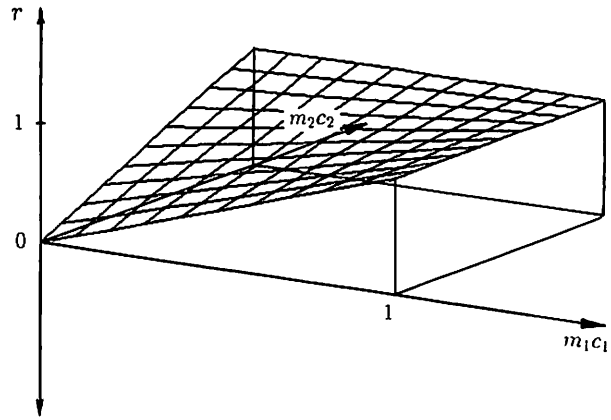


Figure 4: Soft OR mixing function for $K = 2$.

spline curves are present; the multiple curve case formally obeys a WRITE-BLACK imaging model.

4 Objective Function and Learning Procedure

The learning procedure follows the standard two-phase paradigm employed by the EM algorithm and others of its ilk. Both learning and measurement (computing hidden unit activities encoding a data point) operate in the context of an objective function that evaluates prediction errors. Log-likelihood is a suitable choice, where for WRITE-BLACK data sets, 0 represents an OFF data value and 1 represents ON. The objective function for a single data point is

$$_{wb}g_i = \sum_j \log \left[d_{i,j} r_{i,j} + (1 - d_{i,j})(1 - r_{i,j}) \right] \tag{4.1}$$

According to equation 3.1, the predictions $r_{i,j}$ are functionally dependent on the vector of hidden unit responses m_i. These are chosen to be those that optimize the predictions, that is, that maximize g_i. Unfortunately these responses cannot effectively be computed in closed form, and must be solved for by gradient ascent. Figure 6 offers a simple illustration that optimal responses m cannot be computed independently per hidden unit, but instead are interdependent. We have found that attempts to compute hidden unit responses by one-pass feedforward activation rules provide poor enough estimates of the optimum that the learning parameters c become unable to track the correct gradient accurately and fail to discover

underlying multiple cause structure in test data; it is necessary to have the optimal m, computed iteratively. The objective landscape appears to be convex in m however,[1] and we have found that in practice the optimum can be reached from a starting point of $m_k = 0.5$ usually in fewer than five iterations using a conjugate gradient method.

As for learning, the global objective function for an entire training set of I data points is

$$G = \sum_i g_i \tag{4.2}$$

The weights $c_{j,k}$ are found through gradient ascent in G. Note that the gradient

$$\frac{\partial G}{\partial c} = \sum_i \frac{\partial g_i}{\partial c}$$

is functionally dependent on the hidden responses m_i, which differ from data point to data point. These must be updated at each training step. Thus the two-phase computation resembles Boltzmann Machine training, with hidden unit response searches occurring within an overall weight space search (Ackley *et al.* 1985).

5 WRITE-WHITE-AND-BLACK Data Domains

The imaging model and voting rule perspectives on mixing functions suggest that multiple cause domains might exist that are well modeled by modes of causal interaction other than the disjunctive form discussed above. For example, what if hidden units are allowed not only to either abstain or vote some degree of "yes" toward a prediction unit's activity being ON, but also to vote "no," that it should be turned OFF? This amounts to permitting both positive and negative connection weights.

Such an interpretation applies to the data set of Figure 7. These data reflect two independent processes, one of which controls the positions of the black and white squares on the left-hand side, the other controlling the right. A perspicuous multiple cause representation for these data is shown in Figure 10b, consisting of six hidden cluster-centers, three pertaining to the left-hand side, the other three pertaining to the right. This data set reflects a WRITE-WHITE-AND-BLACK imaging model because the hidden causes are responsible for driving both white (OFF) and black (ON) predictions. Gray levels indicate dimensions for which a cluster-center adopts a "don't-know/don't-care" assertion, leaving those pixels to be colored by some other hidden unit(s).

[1] It can be shown through differentiation of 4.1 that for every k', the gradient $\partial g/\partial m_{k'}$ contains at most one local minimum on the interval $0 \leq m_{k'} \leq 1$ for all fixed activation values of the remaining hidden units m_k: $k \neq k'$. This strongly suggests convexity, but leaves open the remote possibility of multiple local minima separated by pathological saddle points. Thus far in our investigations we have observed only convex objective function surfaces.

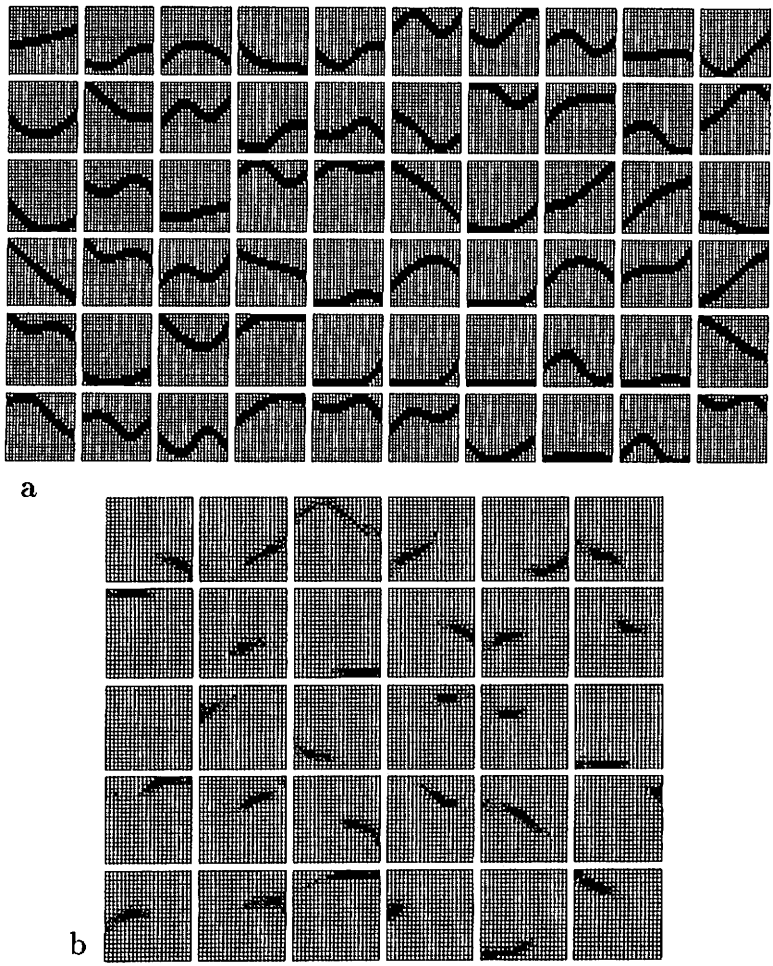


Figure 5: (a) Data samples consisting of randomly generated spline curves. (b) 30-component multiple cause representation for (a) discovered using a soft OR mixing function. 500 training samples were used. *Continued facing page.*

Although a variety of candidate voting schemes are available for modeling the interaction of hidden causes in WRITE-WHITE-AND-BLACK data domains, not all lead to the discovery of independent componential structure as reflected in the left- and right-hand sides of the Figure 7 test data. For example, one possible voting scheme is linear summation,

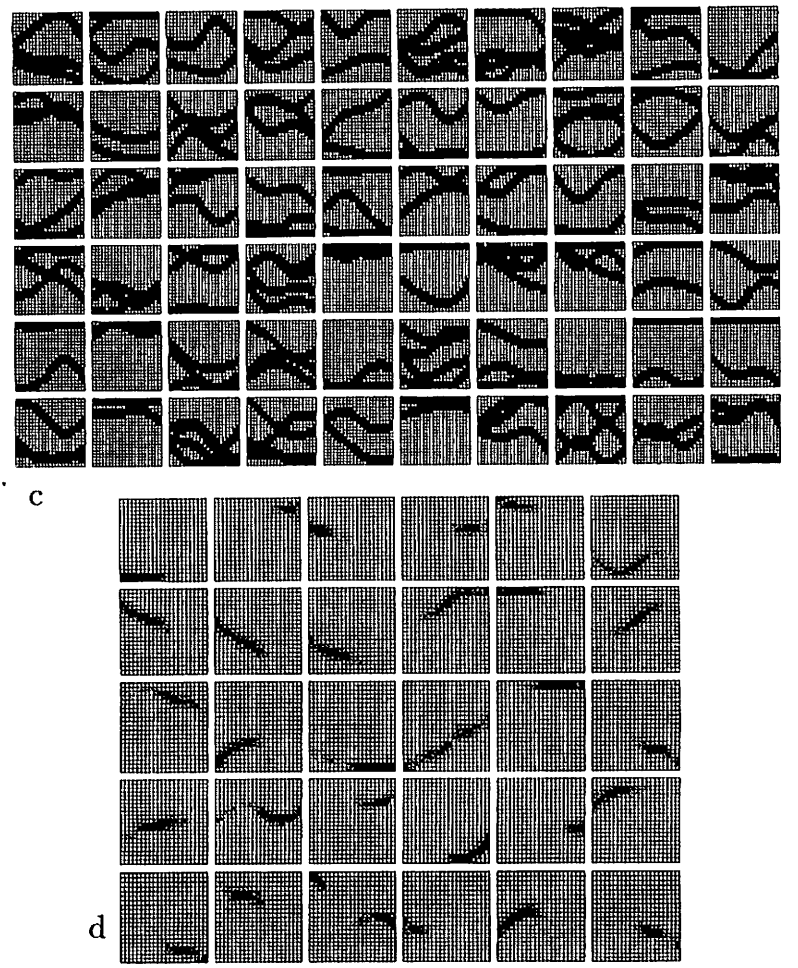


Figure 5: (c) Data samples consisting of several randomly generated spline curves written disjunctively into the data space. (d) 30-component multiple cause representation discovered for (c) using a soft OR mixing function to reflect the WRITE-BLACK disjunctive imaging model. 500 training samples were used.

as employed by principal components analysis. The principal components representation for the Figure 7 data is shown in Figure 8. Principal components is able to reconstruct the data without error using only four hidden units (plus fixed centroid), but these vectors obscure the composi-

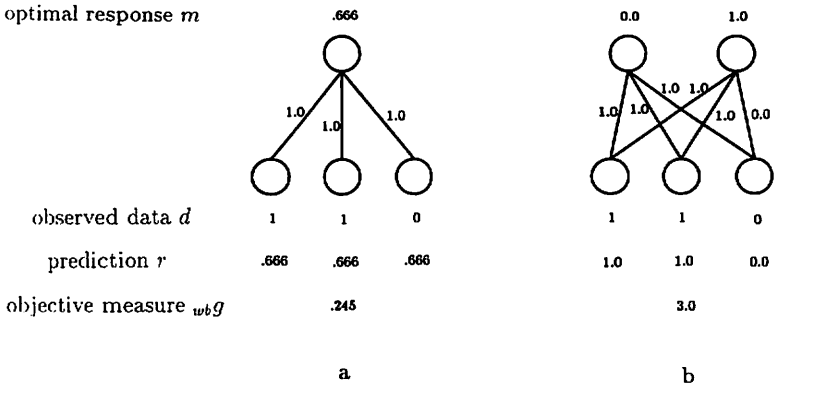


Figure 6: Illustration of the interdependencies of optimal encoder layer responses. Predictions r_j are computed using the optimal activities m shown under the WRITE-BLACK mixing function. (a) The single cluster-center [1 1 1] cannot afford to respond fully to the data vector [1 1 0] because by equation 4.1 the incorrect prediction of a 1 on $d_3 = 0$ would be very costly in terms of the objective function. Instead, the compromise response of 0.66 is optimal. (b) When a second cluster-center [1 1 0] is introduced, it accounts for the observed data by responding fully, leaving the first cluster to adjust its activity to 0 which removes error in the prediction of d_3. Thus if hidden units are regarded as feature detectors, their sensitivity to presented patterns depends upon the context of the other feature detectors available to account for the data observed.

tional structure of the data in that they reveal nothing about the statistical independence of the left- and right-hand processes. Similar results obtain for multiple cause unsupervised learning using a Harmonium network and for a feedforward network using the sigmoid nonlinearity.

By linearly summing hidden unit activities as a first step in the activation function, principal components and most neural net formulations permit errors in predictions by some hidden units to be directly cancelled out by correct predictions from others—without consequence in terms of error in the net prediction. As a result, there is little global pressure for cluster-centers to adopt don't-know values when they are not quite confident about their predictions, and the result is the kind of incoherent representation witnessed in Figure 8. This problem occurs whether or not a sigmoid or other nonlinearity is performed after the summation step.

Instead, a multiple cause formulation delivering coherent cluster-centers requires a different form of nonlinearity in the mixing function. Instead of being able to sum linearly so that a full ON prediction can

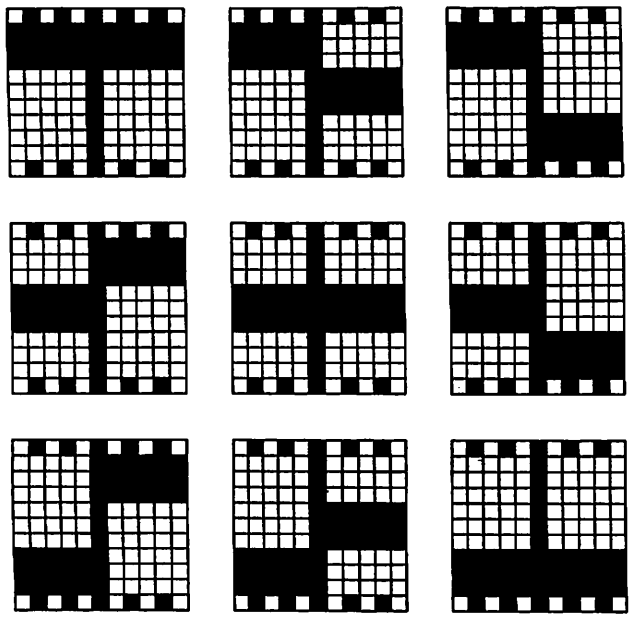


Figure 7: Nine 121-dimensional test data samples exhibiting multiple cause structure. Independent processes control the position of the black rectangle on the left- and right-hand sides.

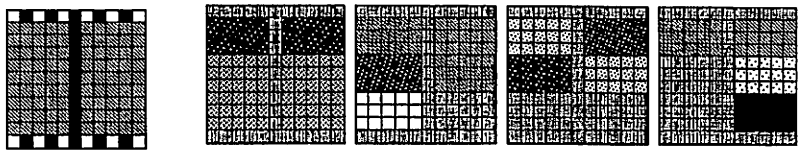
be made if the number of cluster-centers voting ON simply outnumbers those voting OFF and vice versa, active disagreement must result in a net UNCERTAIN or neutral prediction that results in nonzero error when compared with observed data.

The following formalism achieves this purpose. First, let us define the representation of activity and its interpretation for WRITE-WHITE-AND-BLACK data domains.

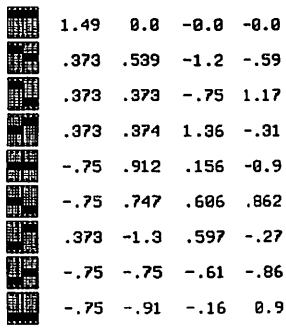
At the data layer ON $\equiv 1$ and OFF $\equiv -1$; at the encoding layer, NULL RESPONSE $\equiv 0$; MAXIMAL RESPONSE $\equiv 1$:

observed data: $d_{i,j} \in \{-1, 1\}$
weights: $-1 \leq c_{j,k} \leq 1$
predictions: $-1 \leq r_{i,j} \leq 1$
measurements: $0 \leq m_{i,k} \leq 1$

We employ a *zero-based representation* at the data layer because it simplifies the subsequent mathematical expressions. The sign of a weight



a b



1.49	0.0	-0.0	-0.0
.373	.539	-1.2	-.59
.373	.373	-.75	1.17
.373	.374	1.36	-.31
-.75	.912	.156	-0.9
-.75	.747	.606	.862
.373	-1.3	.597	-.27
-.75	-.75	-.61	-.86
-.75	-.91	-.16	0.9

c

Figure 8: Principal components representation for the test data from Figure 7. (a) Centroid (white: -1, black: 1). (b) Four component vectors sufficient to encode the nine data points. (lighter shadings: $c_{j,k} < 0$; gray: $c_{j,k} = 0$; darker shadings: $c_{j,k} > 0$). (c) Activities m_i (projections) for the principal components representation of each of the nine test data points.

$c_{j,k}$ indicates whether activity in cluster-center k predicts a 1 or -1 at data dimension j, and its magnitude indicates strength of belief; $c_{j,k} = 0$ corresponds to "don't-know/don't-care" (gray in Fig. 10b). Under the zero-based representation, a convenient form of the log-likelihood objective function evaluating prediction errors becomes,

$$_{wwb}g_{i,j} = \log_2(1 + d_{i,j}r_{i,j}) \tag{5.1}$$

A mixing function achieving the desired form of nonlinearity is constructed such that the opinion of cluster-center C_k about the prediction

activity on the jth data dimension is given by the product, $m_{i,k}c_{j,k}$. The sign of this quantity signifies preference for OFF versus ON, while the magnitude indicates degree of conviction. The voting rule for combining beliefs from several cluster-centers is designed such that a great deal of belief that $r_{i,j}$ should be 1 must outweigh a lesser amount of belief that $r_{i,j}$ should be -1 and vice versa, while roughly equal amounts of belief in each must result in deadlock ($r_{i,j} \approx 0$) as discussed above. Furthermore, the degree of influence any cluster-center has on the outcome decreases as its conviction $|m_{i,k}c_{j,k}|$ approaches 0 ("don't care").

These criteria may be achieved by specifying the way in which positive and negative beliefs balance one another in boundary cases where the beliefs take extreme values $m_{i,k}c_{j,k} \in \{-1, 0, 1\}$ and then assuming bilinear interpolation between these extremes. A satisfactory boundary condition is simply a normalized weighted sum of positive and negative influences:

$$r_{i,j} = \begin{cases} \frac{\sum_k m_{i,k}c_{j,k}}{\sum_k m_{i,k}|c_{j,k}|} & \{\forall c_{j,k}, m_{i,k} : m_{i,k}c_{j,k} \in \{-1, 0, 1\}\} \\ 0 & \sum_k m_{i,k}|c_{j,k}| = 0 \end{cases} \tag{5.2}$$

These boundary conditions specify values at the corners of 2^K K-dimensional hypercubes packed about the origin as illustrated in Figure 9. Note that when any activity $m_{i,k} = 0$, that cluster-center drops out from having any influence on the predictions $r_{i,j}$ and the effective dimensionality of the hypercube decreases by 1. Due to the denominator, conflicting predictions arising from active $c_{j,k}$s of opposite sign end up driving the prediction toward a noncommital 0.

Bilinear interpolation is exponentially expensive in the dimension K, so computation of this mixing function is prohibitively expensive for any sizable number of active cluster-centers. We can, however, offer a computationally tractable approximation to the ideal mixing function. Namely, take as the composite prediction $r_{i,j}$ the quantity

$$_{\text{WAVE}}r_{i,j} = \tag{5.3}$$

$$\frac{\left[\sum_{k:c_{j,k}<0} m_{i,k}(-c_{j,k})\right]\left[\prod_{k:c_{j,k}<0}(1 + m_{i,k}c_{j,k}) - 1\right] + \left[\sum_{k:c_{j,k}>0} m_{i,k}c_{j,k}\right]\left[1 - \prod_{k:c_{j,k}>0}(1 - m_{i,k}c_{j,k})\right]}{\sum_k m_{i,k}|c_{j,k}|}$$

Measurement and learning are performed as described in Section 4. Note, however, that special care must be taken in implementation of the gradient ascent algorithm because the gradient of the WRITE-WHITE-AND-BLACK mixing function becomes discontinuous at $c_{j,k} = 0$, as terms shift between the $c_{j,k} < 0$ and $c_{j,k} > 0$ portions of the numerator of equation 5.3. This reflects the expected qualitative difference between combining beliefs that agree in sign versus those that disagree.

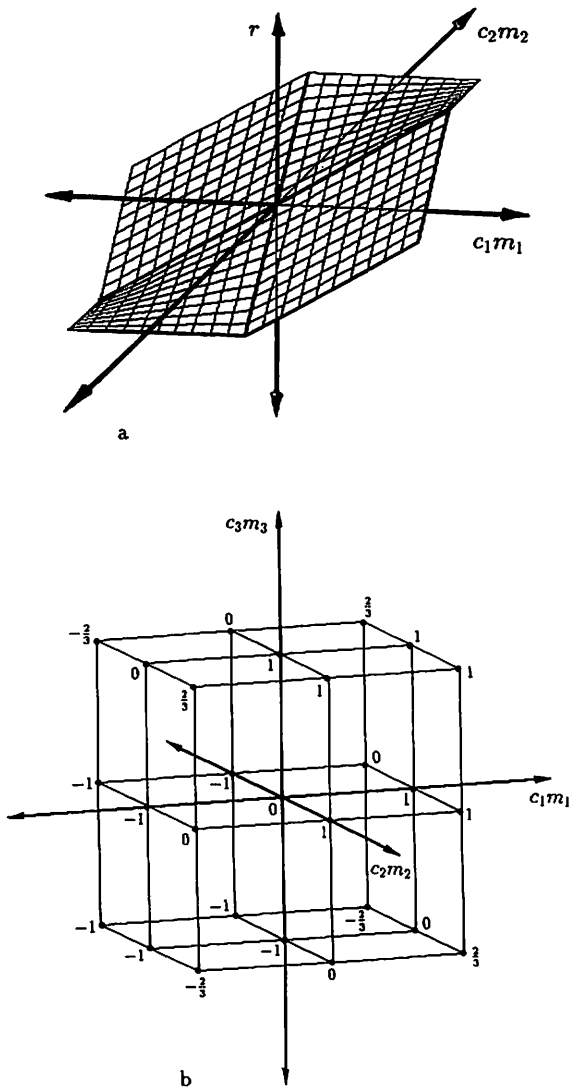


Figure 9: Ideal WRITE-WHITE-AND-BLACK mixing function. (a) Interpolating bilinear surface $r_{i,j}$ as a function of $m_{i,k}c_{j,k}$ for $K = 2$. (b) Boundary values for $r_{i,j}$ defined at the corners of the eight hypercubes for $K = 3$.

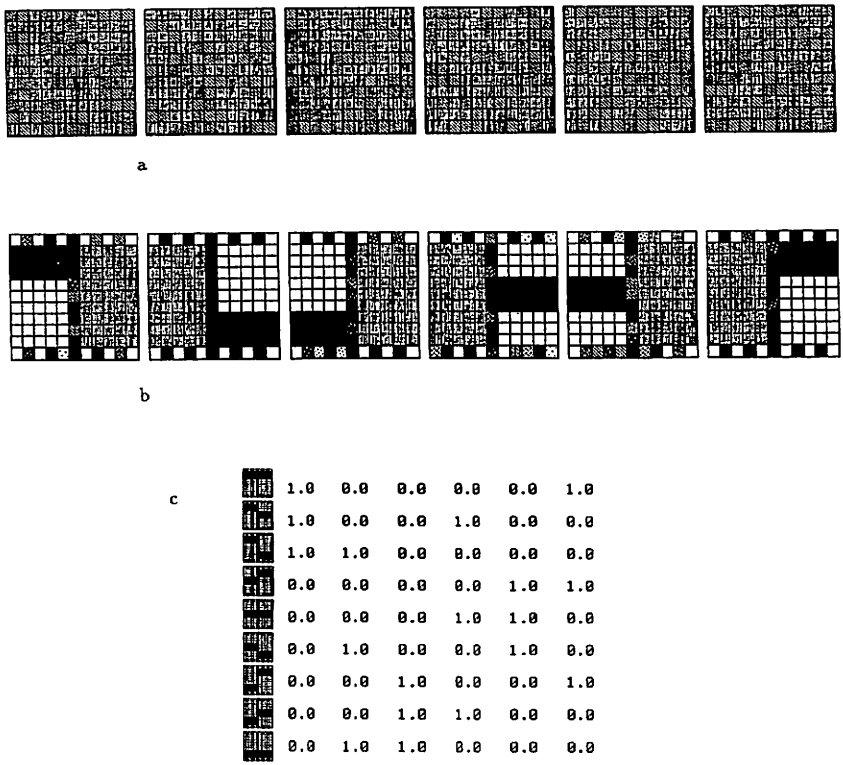


Figure 10: Multiple Cause Mixture Model WRITE-WHITE-AND-BLACK representation for the test data from Figure 7. (a) Initial random cluster-centers. (b) Cluster-centers after seven training iterations (white: $c_{j,k} = -1$; gray: $c_{j,k} = 0$; black: $c_{j,k} = 1$). (c) Activities m_i of the six cluster centers of (b) for the nine training data points. This representation predicts the test data set without error for an objective measure $G = 1089.0$.

6 Experiments

Figure 10 shows that the WRITE-WHITE-AND-BLACK mixing function of equation 5.3 leads to convergence to the coherent multiple cause representation for the test data of Figure 7 starting with random initial weights. The model is robust with respect to noisy training data as indicated in Figure 11.

Although training can be performed for any number K of random

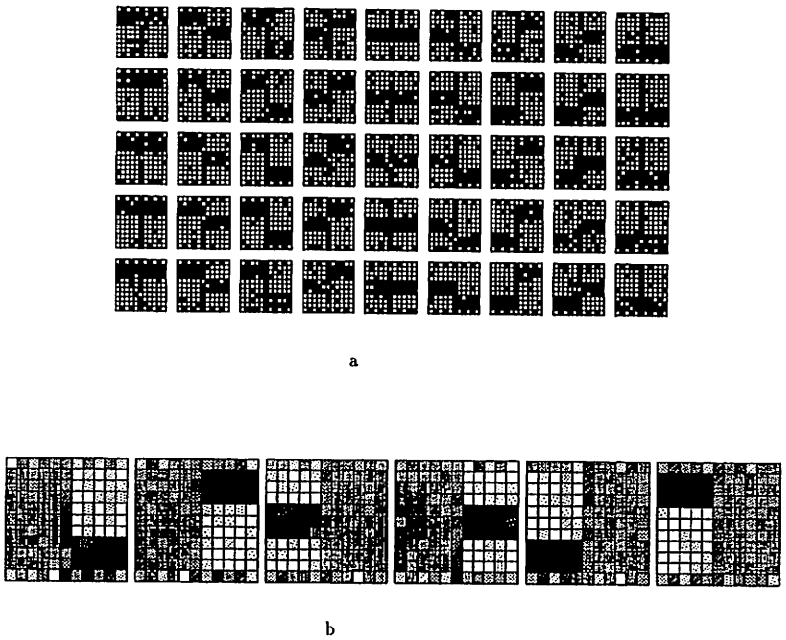


Figure 11: Multiple Cause Mixture Model results for noisy training data. (a) Five test data sample suites with 10% bit-flip noise. Twenty suites were used to train from random initial cluster-centers, resulting in the representation shown in (b). *Continued facing page.*

initial cluster-centers, robustness with respect to local minima in weight space is enhanced by building the model incrementally, starting with K = 1 and adding cluster-centers one at a time until the desired target K is reached. At each step, an evaluation is performed to determine which cluster-center is most responsible for prediction error, and this is split and each child cluster-center slightly perturbed to break symmetry.

This method was used in training the model on data consisting of 21 × 21 pixel images of registered lower-case characters. Results for K = 14 are shown in Figure 12 indicating that the model has discovered statistical regularities associated with ascenders, descenders, circles, etc. Figure 12c shows the hidden feature responses to several noisy versions of the data, and their reconstructions from these components.

Due to the optimization basis for the measurement function, meaningful responses can be computed for incomplete data (Ahmad and Tresp 1993). Missing data are represented by $d_{i,j} = 0$; by equation 5.1 the corresponding prediction $r_{i,j}$ may then float freely without affecting the

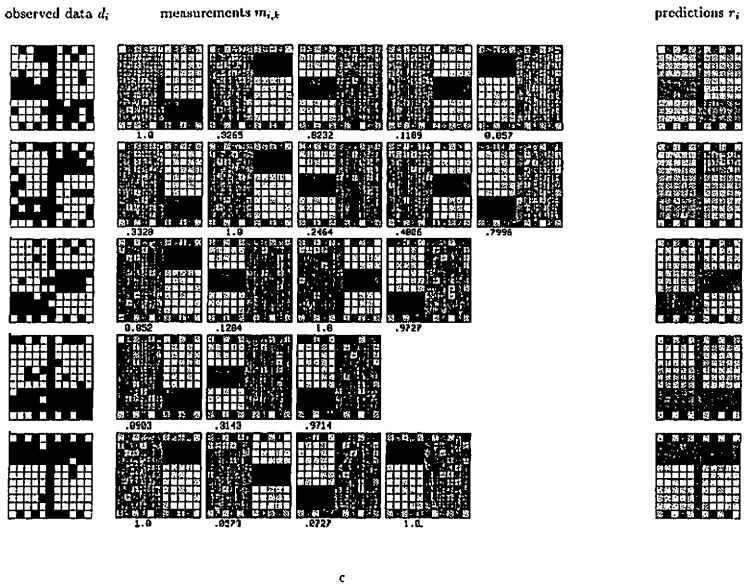
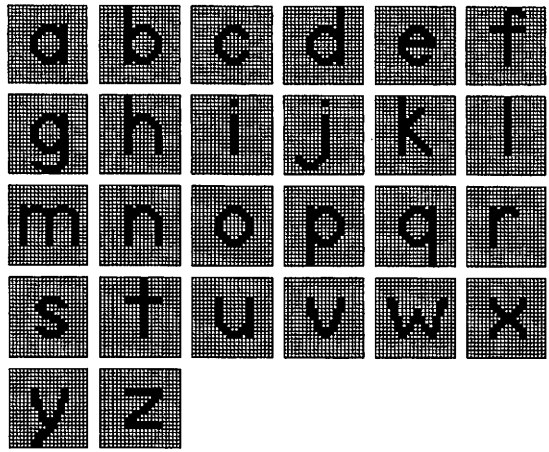


Figure 11: (c) Left: Five test data samples d_i; Middle: Numerical activities $m_{i,k}$ for the most active cluster-centers (the corresponding cluster-center is displayed above each $m_{i,k}$ value); Right: reconstructions (predictions) r_i based on the activities. Note how these "clean up" the noisy samples from which they were computed.

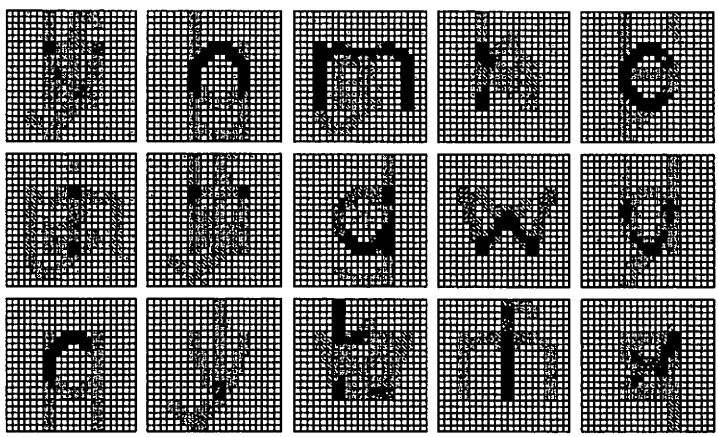
objective function. Figure 13 illustrates reconstructions of noisy *and* incomplete data in the two-process test case.

7 Conclusion

Whether termed "multiple cause," "componential," or "factorial," the significance of this distributed type of representation is suggested by the multiplicity of ways in which high-dimensional observed data may arise from independent processes, each of which pertains only to subspaces of the full observation space. For unsupervised learning algorithms, the difficulty lies in getting the internal knowledge-bearing entities sensibly to divvy up responsibility for training data not just pointwise, but dimensionwise. Instead of attempting to achieve certain statistical properties such as sparseness (Földiák 1990; Hinton and Zemel 1994) or independence of hidden unit responses (Barlow 1989; Schmidhuber 1992), this paper shifts focus to the modes of interaction among hidden causes. We



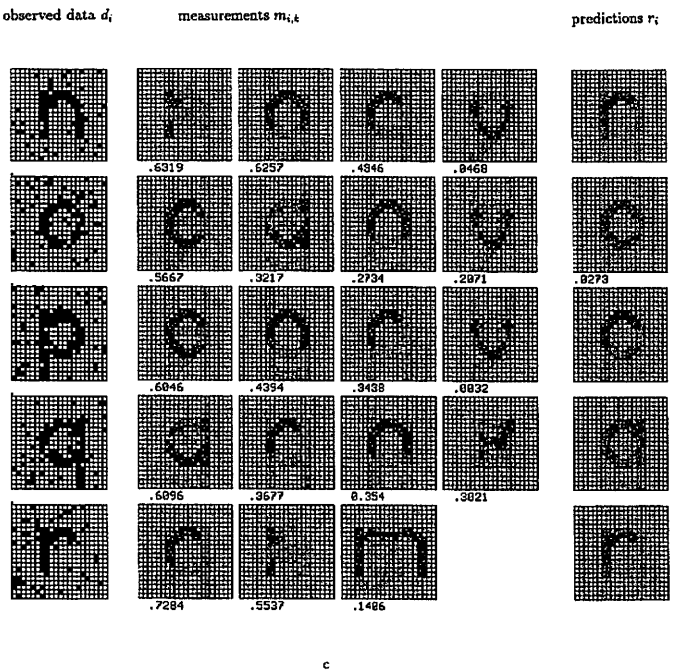
a



b

Figure 12: (a) Training set of 26 441-dimensional binary vectors. (b) Multiple Cause Mixture Model representation at $K = 15$. *Continued facing page.*

have distinguished two different types of multiple cause binary data domain and have shown that appropriately tuned mixing functions—quite different from the standard linear sum followed by sigmoid squashing—permit recovery of the component cluster features. The metaphors of imaging models and voting rules provide conceptual support in design-



c

Figure 12: (c) Left: Five test data samples d_i corrupted with 10% bit-flip noise; Middle: Numerical activities $m_{i,k}$ for the most active cluster-centers (the corresponding cluster-center is displayed above each $m_{i,k}$ value); Right: reconstructions (predictions) r_i based on the activities. Note: to encode this noisy data the cluster-centers discovered on clean data and shown in (b) were clipped to $-0.9 \leq c_{j,k} \leq 0.9$.

ing mixing functions with appropriate functional behaviors. Obviously the notion of tuning mixing functions to the data source can be extended, for example, to continuous valued data. The appropriate representation and treatment of "don't know/don't care" beliefs stand as a key issue in this endeavor.

References

Ackley, D., Hinton, G., and Sejnowski, T. 1985. A learning algorithm for Boltzmann machines. *Cog. Sci.* 9, 147–169.

Ahmad, S., and Tresp, V. 1993. Some solutions to the missing feature problem in vision. In *Advances in Neural Information Processing Systems 5*, S. Hanson,

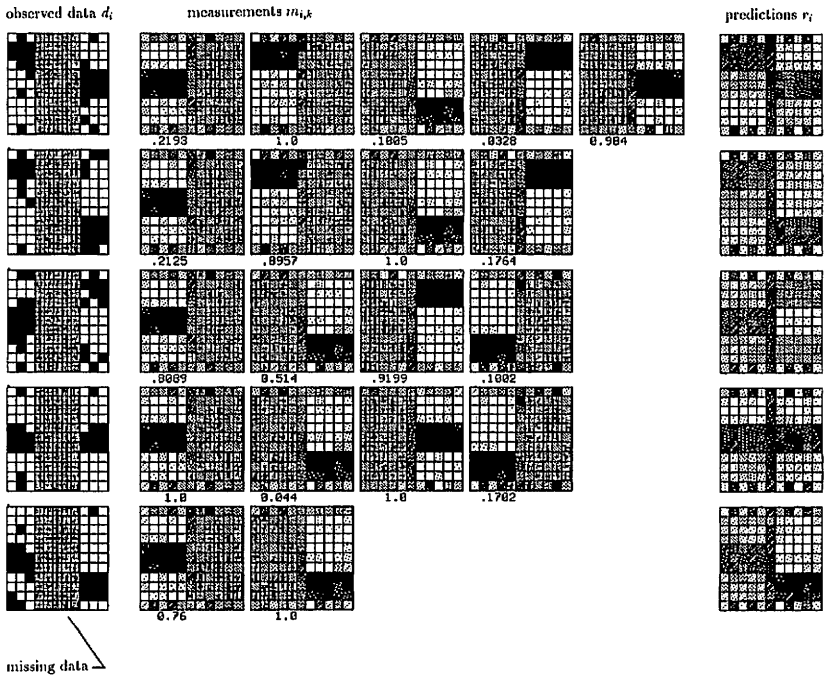


Figure 13: Responses and predictions of a Multiple Cause Mixture Model, trained on noisy data, to incomplete as well as noisy data. Missing data are denoted by gray entries in the "observed data" column.

J. Cowan, and C. Giles, eds., pp. 393–400. Morgan Kaufmann, San Mateo, CA.

Barlow, H. 1989. Unsupervised learning. *Neural Comp.* **1**, 295–311.

Bourlard, H., and Kamp, Y. 1988. Auto-association by multilayer perceptrons and singular value decomposition. *Biol. Cybernet.* **59**(4–5), 291–294.

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

Földiák, P. 1990. Forming sparse representations by local anti-Hebbian learning. *Biol. Cybernet.* **64**(2), 165–170.

Freund, Y., and Haussler, D. 1992. Unsupervised learning of distributions on binary vectors using two-layer networks. In *Advances in Neural Information Processing Systems 4*, J. Moody, S. Hanson, and R. Lippman, eds., pp. 912–919. Morgan Kaufmann, San Mateo, CA.

Hinton, G., and Zemel, R. 1994. Autoencoders, minimum description length and Helmholtz free energy. In *Advances in Neural Information Processing Systems 6*, J. Cowan, G. Tesauro, and J. Alspector, eds., pp. 3–10. Morgan Kaufmann, San Mateo, CA.

Nowlan, S. 1990. Maximum likelihood competitive learning. In *Advances in Neu-*

ral Information Processing Systems 2, D. Touretzky, ed., pp. 574–582. Morgan Kaufmann, San Mateo, CA.

Sanger, T. 1989. An optimality principle for unsupervised learning. In *Advances in Neural Information Processing Systems*, D. Touretzky, ed., pp. 11–19. Morgan Kaufmann, San Mateo, CA.

Schmidhuber, J. 1992. Learning factorial codes by predictability minimization. *Neural Comp.* **4**, 863–879.

Warnock, J., and Wyatt, D. 1982. A device independent graphics imaging model for use with raster devices. *Proc. ACM SIGGRAPH*, pp. 313–319.

Zemel, R. 1993. A minimum description length framework for unsupervised learning. Ph.D. Thesis, Department of Computer Science, University of Toronto.

Received June 29, 1993; accepted April 11, 1994.