DIFFUSION NETWORKS, PRODUCT OF EXPERTS, AND FACTOR ANALYSIS

Tim K. Marks & Javier R. Movellan

University of California San Diego
9500 Gilman Drive
La Jolla, CA 92093-0515

ABSTRACT

Hinton (in press) recently proposed a learning algorithm called contrastive divergence learning for a class of probabilistic models called product of experts (PoE). Whereas in standard mixture models the "beliefs" of individual experts are averaged, in PoEs the "beliefs" are multiplied together and then renormalized. One advantage of this approach is that the combined beliefs can be much sharper than the individual beliefs of each expert. It has been shown that a restricted version of the Boltzmann machine, in which there are no lateral connections between hidden units or between observation units, is a PoE. In this paper we generalize these results to diffusion networks, a continuous-time, continuous-state version of the Boltzmann machine. We show that when the unit activation functions are linear, this PoE architecture is equivalent to a factor analyzer. This result suggests novel non-linear generalizations of factor analysis and independent component analysis that could be implemented using interactive neural circuitry.

1. INTRODUCTION

Hinton (in press) recently proposed a fast learning algorithm, known as contrastive divergence learning, for a class of models called product of experts (PoE). In standard mixture models the "beliefs" of individual experts are averaged, but in PoEs the "beliefs" are multiplied together and then renormalized, i.e.,

$$p(o) \propto \prod_j p_j(o) \tag{1}$$

where $p_j(o)$ is the probability of the observed vector o under expert j and $p(o)$ is the probability assiged to o when the opinions of all the experts are combined. One advantage of this approach is that the combined beliefs are sharper than the individual beliefs of each expert, potentially avoiding the problems of standard mixture models when applied to high-dimensional problems. It can be shown that a restricted version of the Boltzmann machine, in which hidden units and observable units form a bipartite graph, is in fact a PoE (Hinton, in press). Indeed, Hinton (in press) recently trained these restricted Boltzmann machines (RBMs) using contrastive divergence learning with very good results. One problem with Boltzmann machines is that they use binary-state units, a representation which may not be optimal for continuous data such as images and sounds.

This paper concerns the diffusion network, a continuous-time, continuous-state version of the Boltzmann machine. Like Boltzmman machines, diffusion networks have bidirectional connections, which in this paper we assume to be symmetric. If a diffusion network is given the same connectivity structure as an RBM, the result will also be a PoE model.

The main result presented here is that when the activation functions are linear, these restricted diffusion networks model the same class of observable distributions as factor analysis. This is somewhat surprising given the fact that diffusion networks are feedback models while factor analyzers are feedforward models. Most importantly, the result suggests novel non-linear generalizations of factor analysis and independent component analysis that could be implemented using interactive circuitry. We show results of simulations in which we trained restricted diffusion networks using contrastive divergence.

2. DIFFUSION NETWORKS

A diffusion network is a neural network specified by a stochastic differential equation

$$dX(t) = \mu(X(t))dt + \sigma dB(t) \tag{2}$$

where $X(t)$ is a random vector describing the state of the system at time t, and B is a standard Brownian motion process. The term σ is a constant, known as the dispersion, that controls the amount of noise in the system. The function μ, known as the drift, represents the deterministic kernel of the system. If the drift is the gradient of an energy function, $\mu(x) = -\nabla_x \phi(x)$, and the energy function satisfies certain conditions (Gidas, 1986), then it can be shown that X has a limit probability density $p(x) \propto \exp(-2\,\phi(x)/\sigma^2)$ which is independent of the initial conditions.

2.1. Linear diffusions

If we let $\sigma = \sqrt{2}$ and the energy is a quadratic function $\phi(x) = (1/2)x^T ax$, where a is a symmetric, positive definite matrix, then the limit distribution is

$$p(x) \propto \exp(-\frac{1}{2}x^T ax). \tag{3}$$

This distribution is Gaussian with zero mean and covariance matrix a^{-1}. Taking the gradient of the energy function, we get

$$\mu(x) = -\nabla_x \phi(x) = -ax \tag{4}$$

and thus the activation dynamics are linear:

$$dX(t) = -aX(t)dt + \sigma dB(t) \tag{5}$$

These dynamics have the following neural network interpretation: Let $a = (d - w)$, where w represents a matrix of synaptic conductances and d is a diagonal matrix of transmembrane conductances d_i (current leakage terms). We then get the following form

$$dX_i(t) = [\bar{X}_i(t) - d_i X_i(t)]dt + \sigma dB(t) \tag{6}$$

where

$$\bar{X}_i(t) = \sum_j w_{ij} X_j(t) \tag{7}$$

is interpreted as the net input current to neuron i and $X_i(t)$ as the activation of that neuron.

3. FACTOR ANALYSIS

Factor analysis is a probabilistic model of the form

$$\tilde{O} = c\tilde{H} + Z \tag{8}$$

where $\tilde{O}$ is an n_o dimensional random vector representing observable data; $\tilde{H}$ is an n_h dimensional Gaussian vector representing hidden independent sources, $\mathrm{E}(\tilde{H}) = 0$, $\mathrm{Cov}(\tilde{H}) = I$, the identity matrix; and Z is a Gaussian vector independent of $\tilde{H}$ representing noise in the observation generation process, $\mathrm{E}(Z) = 0$, $\mathrm{Cov}(Z) = \Psi$, a diagonal matrix. If $\tilde{H}$ is logistic instead of Gaussian, and in the limit as $\Psi \to 0$, Equation (8) is the standard model underlying independent component analysis (ICA) (Bell & Sejnowski, 1995; Attias, 1999). Note that in factor analysis, as well as in ICA, the observations are generated in a feedforward manner from the hidden sources and the noise process.

It follows from Equation 8 that $\tilde{O}$ and $\tilde{H}$ have covariances

$$\mathrm{Cov}(\tilde{X}) = \mathrm{Cov}\begin{pmatrix}\tilde{O}\\ \tilde{H}\end{pmatrix} = \begin{bmatrix} cc^T + \Psi & c \\ c^T & I \end{bmatrix} \tag{9}$$

Taking the inverse of this block matrix, we get

$$(\mathrm{Cov}(\tilde{X}))^{-1} = \begin{bmatrix} \Psi^{-1} & -\Psi^{-1}c \\ -c^T\Psi^{-1} & I + c^T\Psi^{-1}c \end{bmatrix} \tag{10}$$

4. RESTRICTED DIFFUSION NETWORKS

In this section, we discuss a subclass of diffusion networks that have a PoE architecture. To begin with, we divide the state vector of a diffusion network into observable and hidden units, $X^T = (O^T, H^T)$. It is easy to show that if the connectivity matrix is restricted so that there are no lateral connections from hidden units to hidden units and no lateral connections from observable units to observable units, i.e., if

$$a = \begin{bmatrix} d_o & -w_{oh} \\ -w_{oh}^T & d_h \end{bmatrix} \tag{11}$$

where d_o and d_h are diagonal, then the limit distribution of X is a PoE. Hereafter, we refer to a diffusion network with no hidden-hidden or observable-observable connections as a restricted diffusion network (RDN).

Suppose we are given an arbitrary linear RDN. The limit distribution, $p(OH)$, has Gaussian marginal distributions $p(O)$ and $p(H)$ whose covariances can be found by inverting the block matrix in Equation 11:

$$\mathrm{Cov}(H) = (d_h - w_{oh}^T d_o^{-1} w_{oh})^{-1} \tag{12}$$

$$\mathrm{Cov}(O) = (d_o - w_{oh} d_h^{-1} w_{oh}^T)^{-1} \tag{13}$$

Let $\mathcal{D}^o$ represent the set containing every distribution on $\mathbb{R}^{n_o}$ that is the limit distribution of the observable units of any linear RDN with n_h hidden units. In some cases, we can think of the hidden units of a network as the network's internal representation of the world. Some authors, such as Barlow (1994), have postulated that one of the organizing principles of early processing in the brain is for these internal representations to be as independent as possible. Thus it is of interest to study the set of linear RDNs whose hidden units are independent, i.e., for which $\mathrm{Cov}(H)$ is diagonal. One might wonder, for example, whether restricting $\mathrm{Cov}(H)$ to be the identity matrix imposes any constraints on $\mathrm{Cov}(O)$.

Let $\mathcal{D}_i^o$ represent the set containing every distribution on R^{n_o} that is the limit distribution of the observable units of any linear RDN with n_h hidden units that are independent, $\mathrm{Cov}(H) = I$. We now show that forcing the hidden units to be independent does not constrain the class of observable distributions that can be represented by a linear RDN.

Theorem 1 : $\mathcal{D}_i^o = \mathcal{D}^o$.

Proof : Since $\mathcal{D}_i^o$ is a subset of $\mathcal{D}^o$, we just need to show that any distribution in $\mathcal{D}^o$ is also in $\mathcal{D}_i^o$. Let $p(OH)$ be the limit distribution of an arbitrary linear RDN with parameters w_{oh}, d_h, and d_o. We will show there exists a new linear RDN, with limit distribution $p(O'H')$ and parameters w'_{oh}, d'_h, and d_o, such that $\mathrm{Cov}(H') = I$ and $p(O') = p(O)$.

First we take the eigenvalue decomposition

$$I - d_h^{-1/2} w_{oh}^T d_o^{-1} w_{oh} d_h^{-1/2} = \sigma \Lambda \sigma^T \tag{14}$$

where σ is a unitary matrix and Λ is diagonal. Equation 14 can be rewritten

$$w_{oh}^T d_o^{-1} w_{oh} = d_h - d_h^{1/2} \sigma \Lambda \sigma^T d_h^{1/2}. \tag{15}$$

We define

$$d'_h = \Lambda^{-1} \qquad \text{and} \qquad w'_{oh} = w_{oh} d_h^{-1/2} \sigma \Lambda^{-1/2}. \tag{16}$$

By Equation 12 (and after some derivations) we find that

$$\begin{aligned} \mathrm{Cov}(H') &= (d'_h - (w'_{oh})^T d_o^{-1} w'_{oh})^{-1} \\ &= I. \end{aligned} \tag{17}$$

By Equation 13 (and after some derivations) we find that

$$\begin{aligned} \mathrm{Cov}(O') &= (d_o - w'_{oh}(d'_h)^{-1}(w'_{oh})^T)^{-1} \\ &= (d_o - w_{oh} d_h^{-1} w_{oh}^T)^{-1} \\ &= \mathrm{Cov}(O). \qquad \square \end{aligned} \tag{18}$$

Theorem 1 will help to elucidate the relationship between factor analysis and linear RDNs. We will show that the class of distributions over observable variables that can be generated by factor analysis models with n_h hidden sources is the same as the class of distributions over observable units that can be generated by linear RDNs with n_h hidden units.

5. FACTOR ANALYSIS AND DIFFUSION NETS

In this section we examine the relationship between feedforward factor analysis models and feedback diffusion networks. Let $\mathcal{F}^o$ represent the class of probability distributions over observable units that can be generated by factor analysis models with n_h hidden units. In Theorem 2 we will show that this class of distributions is equivalent to $\mathcal{D}^o$, the class of distributions generated by linear RDNs. To do so, we first need to prove a Lemma regarding the subclass of factor analysis models for which the lower right block of the matrix in Equation 10, $I + c^T \Psi^{-1} c$, is diagonal. We define $\mathcal{F}_i^o$ as the class of probability distributions on $\mathbb{R}^{n_o}$ that can be generated by such restricted factor analysis models.

Lemma : $\mathcal{F}_i^o = \mathcal{F}^o$.

Proof : Since $\mathcal{F}_i^o$ is a subset of $\mathcal{F}^o$, we just need to show that any distribution in $\mathcal{F}^o$ is also in $\mathcal{F}_i^o$. Let $p(\tilde{O}\tilde{H})$ be the distribution generated by an arbitrary factor analysis model with parameters c and Ψ. We will show that there exists a new factor analysis model, with joint distribution $p(O'\tilde{H})$ and parameters c' and Ψ, such that $p(O') = p(\tilde{O})$ and such that $I + (c')^T \Psi^{-1} c'$ is diagonal.

First we take the eigenvalue decomposition

$$c^T \Psi^{-1} c = S D S^T \tag{19}$$

where S is a unitary matrix and the eigenvalue matrix D is diagonal. Now define a rotation q by

$$q = S \qquad \text{and} \qquad c' = cq. \tag{20}$$

Then

$$(c')^T \Psi^{-1} c' = D \tag{21}$$

which is diagonal. Thus $I + (c')^T \Psi^{-1} c$ is diagonal. Furthermore, from the equation $O' = c'\tilde{H} + \tilde{Z}$ we can derive $\mathrm{Cov}(O') = c'(c')^T + \Psi = cc^T + \Psi = \mathrm{Cov}(\tilde{O})$. $\square$

Now we are ready to prove the main result of this paper: the fact that factor analysis models are equivalent to linear RDNs.

Theorem 2 : $\mathcal{D}^o = \mathcal{F}^o$.

Proof :
[⇒] : $\mathcal{F}^{\mathbf{o}} \subset \mathcal{D}^{\mathbf{o}}$. Let $p(\tilde{O})$ be the distribution over observable units generated by any factor analysis model. By the Lemma, there exists a factor analysis model with distribution $p(\tilde{X}^T) = p(\tilde{O}^T, \tilde{H}^T)$ and parameters c, Ψ such that its marginal distribution over the observable units equals the given distribution $p(\tilde{O})$ and for which $I + c^T \Psi^{-1} c$ is diagonal.

Let $p(X^T) = p(O^T, H^T)$ be the limit distribution of the linear diffusion network with parameters w_{oh}, d_o, d_h given by setting the connection matrix of Equation 11 equal to the inverse covariance matrix of Equation 10:

$$a = (\mathrm{Cov}(\tilde{X}))^{-1}. \tag{22}$$

Then $d_h = \Psi^{-1}$ and $d_o = I + c^T \Psi^{-1} c$ are both diagonal, so this diffusion net is a RDN. By Equation 3, the limit distribution of X is Gaussian with covariance $\mathrm{Cov}(X) = a^{-1}$, from which it follows that $\mathrm{Cov}(X) = \mathrm{Cov}(\tilde{X})$. In particular, $\mathrm{Cov}(O) = \mathrm{Cov}(\tilde{O})$. Thus, $p(\tilde{O}) = p(O)$ is the limit distribution of a linear RDN.

[⇐] : $\mathcal{D}^{\mathbf{o}} \subset \mathcal{F}^{\mathbf{o}}$. Let $p(O)$ be the limit distribution over observable units of any linear RDN. By Theorem 1, there exists a linear RDN with distribution $p(X^T) = p(O^T, H^T)$ and parameters w_{oh}, d_o, d_h such that its marginal distribution over the observable units equals the given distribution $p(O)$ and for which $\mathrm{Cov}(H) = I$.

Consider the factor analysis model with parameters $\Psi = d_o^{-1}$, $c = d_o^{-1} w_{oh}$. Using reasoning similar to that used in the first part of this proof, one can verify that this factor analysis model generates a distribution over observable units that is equal to $p(O)$.

6. SIMULATIONS

RDNs have continuous states rather than binary states, so they can represent pixel intensities in a more natural manner than restricted Boltzmann machines (RBMs) can. We used contrastive divergence learning rules to train a linear RDN with 10 hidden units in an unsupervised manner on a database of 48 face images (16 people in three different lighting conditions) from the MIT Face Database (Turk & Pentland, 1991). The linear RDN had 3600 observable units (60×60 pixels). Each expert in the product of experts consists of one hidden unit and the learned connection weights between that hidden unit and all of the observable units. These learned connection weights can be thought of as the receptive field of that hidden unit. Figure 1 shows the receptive fields of the 10 hidden units after training. Because of the correspondence between linear RDNs and factor analysis that was proven in previous sections, the receptive field of a hidden unit in the linear RDN should correspond to a column of the factor loading matrix c in Equation 8 (after the column is normalized by premultiplying by Ψ^{-1}, the inverse of the variance of the noise in each observable dimension).

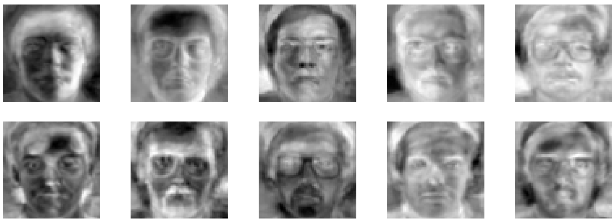

Fig. 1: Receptive fields of the 10 hidden units of a linear RDN that was trained on a database of 48 face images by minimizing contrastive divergence.

Because of the relationship between factor analysis and principal component analysis, one would expect the receptive fields of the hidden units of the linear RDN to resemble the eigenfaces (Turk & Pentland, 1991), which are the eigenvectors of the pixelwise covariance matrix, of the same database. We calculated the eigenfaces of the database. The eigenfaces corresponding to the 10 largest eigenvalues are shown in Figure 2. The qualitative similarity between the linear RDN receptive fields in Figure 1 and the eigenfaces in Figure 2 is readily apparent.

Partially occluded images correspond to splitting the observable units O into those with known values, O_k, and those with unknown values, O_u. Given an image with known values o_k, reconstructing the values of the occluded units involves finding the posterior distribution $p(O_u \mid O_k = o_k)$. The feedback architecture of diffusion networks provides a natural way to find such a posterior distribution: clamp the unoccluded observable units to the values o_k, and let the network settle to equilibrium. The equilibrium distribution of the network will then be the posterior distribution $p(O_u \mid O_k = o_k)$. Using the linear RDN that was trained on face images, we reconstructed occluded versions of the images in the training set by taking the mean of this posterior distribution. Figure 3 shows the results of reconstructing two occluded face images.

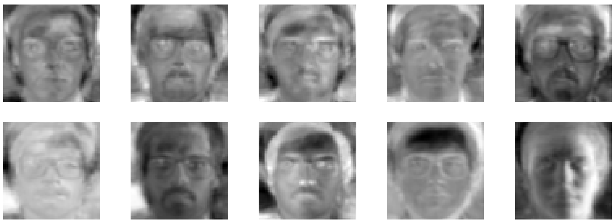

Fig. 2: Eigenfaces calculated from the same database of 48 face images.

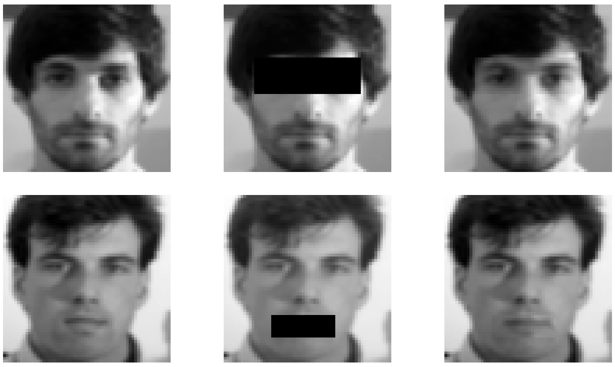

Fig. 3: Reconstruction of two occluded images. *Left:* Image from the training set for the linear RDN. *Center:* The observable units corresponding to the visible pixels were clamped, while the observable units corresponding to the occluded pixels were allowed to run free. *Right:* Reconstructed image shown is the mean of the equilibrium distribution.

7. DISCUSSION

We proved the rather surprising result that a feedback model, the linear RDN, spans the same family of distributions as factor analysis, a feedworward model. Furthermore, we have shown that for this class of feedback models, restricting the marginal distribution of the hidden units to be independent does not restrict the distributions over the observable units that they can generate. Linear RDNs can be trained using the contrastive divergence learning method of Hinton (in press), which is relatively fast. A main advantage of feedback networks over feedforward models is that one can easily solve inference problems, such as pattern completion for occluded images, using the same architecture and algorithm that are used for generation. Most importantly, diffusion networks can also be defined with nonlinear activation functions (Movellan, Mineiro & Williams, In Press), and we have begun to explore nonlinear extensions of the linear case outlined here. Extensions of this work to nonlinear diffusion networks open the door to possibilities of new ICA-like algorithms based on feedback rather than feedforward architectures.

References

Attias, H. (1999). Independent Factor Analysis. *Neural Computation*, *11*(4), 803–851.

Barlow, H. (1994). What is the computational goal of the neocortex? In C. Koch (Ed.), *Large scale neuronal theories of the brain*, pages 1–22. Cambridge, MA: MIT Press.

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

Gidas, B. (1986). Metropolis-type monte carlo simulation algorithms and simulated annealing. In J. L. Snell (Ed.), *Topics in contermporary probability and its applications*, pages 159–232. Boca Raton: CRC Press.

Hinton, G. E. (in press). Training Products of Experts by Minimizing Contrastive Divergence. *Neural Computation*.

Movellan, J. R., Mineiro, P., & Williams, R. J. (In Press). Partially Observable SDE Models for Image sequence recognition tasks. In T. Leen, T. G. Dietterich, & V. Tresp (Eds.), *Advances in Neural Information Processing Systems*, number 13, pages 880–886. Cambridge, Massachusetts: MIT Press.

Turk, M. & Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience, 3(1),* 71–86.