Journal of Machine Learning Research 0 (2003) 0-0 Submitted 0/0; Published 0/0

Energy-Based Models for Sparse Overcomplete Representations

Yee Whye Teh YWTEH@CS.UTORONTO.CA
Max Welling WELLING@CS.UTORONTO.CA
Department of Computer Science, University of Toronto
10 King's College Road, Toronto M5S 3G4, Canada.

Simon Osindero SIMON@GATSBY.UCL.AC.UK
Gatsby Computational Neuroscience Unit, University College London
17 Queen Square, London WC1N 3AR, United Kingdom.

Geoffrey E. Hinton HINTON@CS.UTORONTO.CA
Department of Computer Science, University of Toronto
10 King's College Road, Toronto M5S 3G4, Canada.

Editor: Te-Won Lee, Jean-François Cardoso, Erkki Oja and Shun-Ichi Amari

Abstract

We present a new way of extending independent components analysis (ICA) to overcomplete representations. In contrast to the causal generative extensions of ICA which maintain marginal independence of *sources*, we define *features* as deterministic (linear) functions of the inputs. This assumption results in marginal *dependencies* among the features, but *conditional independence* of the features given the inputs. By assigning energies to the features a probability distribution over the input states is defined through the Boltzmann distribution. Free parameters of this model are trained using the contrastive divergence objective (Hinton, 2002). When the number of features is equal to the number of input dimensions this energy-based model reduces to noiseless ICA and we show experimentally that the proposed learning algorithm is able to perform blind source separation on speech data. In additional experiments we train overcomplete energy-based models to extract features from various standard data-sets containing speech, natural images, hand-written digits and faces.

Keywords: Independent Components Analysis, Density Estimation, Overcomplete Representations, Sparse Representations

1. Introduction

There have been two dominant ways of understanding ICA, one based on a bottom-up, filtering view and the other based on a top-down, causal generative view. In the information maximization approach (Bell and Sejnowski, 1995, Shriki et al., 2002) the aim is to maximize the mutual information between the observations and the non-linearly transformed outputs of a set of linear filters. In the causal generative view (Pearlmutter and Parra, 1996, MacKay, 1996, Cardoso, 1997), on the other hand, the aim is to build a density model in which independent, non-Gaussian sources are linearly combined to produce the observations.

The main point of this paper is to show that there is a third, "energy-based" approach to understanding ICA which combines a bottom-up, filtering view with the goal of fitting a

©2003 Yee Whye Teh, Max Welling, Simon Osindero and Geoffrey E. Hinton.

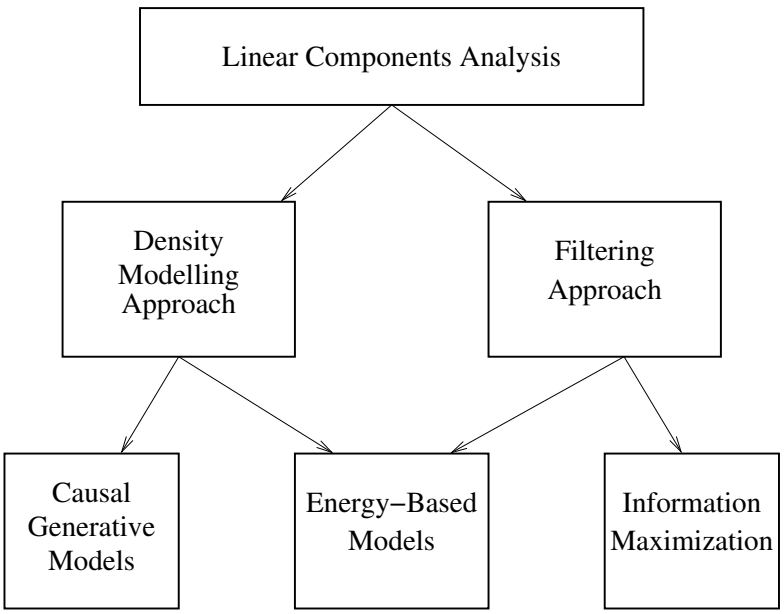

Figure 1: Different methods for non-Gaussian linear components analysis.

probability density to the observations. The parameters of an energy-based model specify a deterministic mapping from an observation vector $\mathbf{x}$ to a feature[1] vector and the feature vector determines a global energy, $E(\mathbf{x})$. The probability density of $\mathbf{x}$ is defined by:

$$p(\mathbf{x}) = \frac{e^{-E(\mathbf{x})}}{Z}$$

where Z is a normalization factor - the integral of the numerator over all possible observation vectors. The energy-based approach is interesting because it suggests a novel and tractable way of extending ICA to overcomplete and multi-layer models.

The relationship between the three approaches is depicted in Figure 1. In general, they are quite different, but they all become equivalent for the "square" and noiseless case, that is, when the number of sources or features equals the number of observations and there is no observation noise. While complete representations have been applied successfully to a wide range of problems, researchers have recently argued for "overcomplete" representations where there are more sources or features than observations. Apart from greater model flexibility, reported advantages include improved robustness in the presence of noise (Simoncelli et al., 1992), more compact and more easily interpretable codes (Mallat and Zhang, 1993) and superresolution (Chen et al., 1998).

The natural way to extend the causal generative approach to the overcomplete case is to retain the assumption that the sources are independent when the model is used to generate data and to accept the consequence that an observation vector creates a posterior distribution over a multiplicity of source vectors. In this posterior distribution, the sources

1. When discussing energy-based models, we use the term "feature" rather than "source" for reasons that will become clear when we discuss extensions to the overcomplete case.

Models	Marginal distribution over 'source' vectors (before an observation)	Conditional distribution over 'source' vectors (after an observation)
Square	Independent (by assumption)	Independent (because deterministic)
Causal overcomplete	Independent (by assumption)	Dependent (explaining-away)
Energy-based overcomplete	Dependent (rejecting-away)	Independent (because deterministic)

Table 1: Independence properties of three types of models.

are conditionally dependent due to the effect known as "explaining-away" and, in general, the distribution has the unfortunate property that it is computationally intractable.

The natural way to extend the information maximization approach to overcomplete representations is to retain both the simple, deterministic, feedforward filtering of observations and the mutual information objective function (Shriki et al., 2002). However, because the manifold of possible filter outputs typically does not consist of the whole space (except in the square case), the equivalence with causal generative models breaks down.

When our energy-based approach to ICA is made overcomplete, it continues to be a proper density model and it retains the computationally convenient property that the features are a deterministic function of the observation vector. However, it abandons the marginal independence of the features (which is why we do not call them sources). A useful way of understanding the difference between energy-based density models and causal generative density models is to compare their independence properties. Table 1 summarizes the similarities and differences.

The table reminds us that the different views are equivalent in the square case, and hence, in the absence of observations, the sources are marginally independent. Further, the posterior distribution over source vectors conditioned on an observation vector collapses to a point in the absence of noise, so the sources are trivially independent in the posterior distribution. In the causal generative approach this conditional independence of the sources is seen as a fortuitous consequence of using as many sources as observations and avoiding noise in the observations, and is not retained in the overcomplete case. In the energy-based view, the conditional independence of the features is treated as a basic assumption that remains true even in the overcomplete case. We can consider the energy contributed by the activity of a feature of an energy-based model as the negative log probability of a one-dimensional, non-Gaussian distribution. However not all combinations of feature activities can occur

because the lower-dimensional observation space only maps to a restricted manifold in the feature space.

The marginal dependence of the features in an overcomplete, energy-based model can be understood by considering an illuminating but infinitely inefficient way of generating unbiased samples from the energy-based density model. First we sample the features independently from their "prior" distributions (the negative exponentials of their individual energy contributions) and then we reject cases in which the feature activities do not correspond to valid observation vectors. This process of "rejecting-away"creates dependencies among the activities of different features.

For some applications, such as unmixing sound sources, the causal generative approach is clearly more appropriate than the energy-based approach because we have a strong prior belief that the sources are marginally independent. In many other applications, however, the real aim is to model the probability density of the data, or to discover interpretable structure in the data, or to extract a representation that is more useful for controlling action than the raw data itself. In these applications, there is no *a priori* reason for preferring the causal generative approach to the energy-based approach that characterizes each observation vector by representing the degree to which it satisfies a set of learned features.

2. Square ICA

In this section we will briefly review standard models for ICA. One of the first expositions on ICA (Comon, 1994) used the entropy of linearly transformed input vectors as a *contrast function* to find statistically independent directions in input space. Indeed many, if not all, ICA algorithms ultimately reduce to optimizing some sort of "contrast function"; this overview will not mention them all. Rather we will focus on reviewing two general approaches to ICA, namely the causal generative approach (Pearlmutter and Parra, 1996, MacKay, 1996, Cardoso, 1997) and the information maximization approach (Bell and Sejnowski, 1995, Shriki et al., 2002). Subsequent sections will then compare these canonical approaches with our proposed energy-based approach, and in particular will explore the consequences of making the different models overcomplete.

Consider a real valued input, denoted by $\mathbf{x}$, of dimensionality D, and an M-dimensional source or feature vector, denoted by $\mathbf{s}$. In this section we will consider the special case where the number of input dimensions is equal to that of the sources or features, i.e. $D = M$.

2.1 The Causal Generative Approach

In the causal generative approach, the sources $\mathbf{s}$ are assumed to be independent, that is, the distribution $p(\mathbf{s})$ factorizes as

$$p(\mathbf{s}) = \prod_{i=1}^{M} p_i(s_i) \tag{1}$$

while the inputs are simply linear combinations of the sources. Moreover, we will assume for now that there is no noise on the inputs, i.e.

$$\mathbf{x} = A\mathbf{s}$$

where A is a square invertible matrix called the mixing matrix. Inverting this relationship we have

$$\mathbf{s} = W\mathbf{x} \qquad W = A^{-1} \tag{2}$$

where the inverse mixing matrix W will be called the filter matrix since each row acts as a linear filter of the inputs.

The aim is now to recover the statistically independent source signals $\mathbf{s}$ from the linearly mixed observations $\mathbf{x}$. This turns out to be possible only if the statistical properties of the sources are non-Gaussian. Thus, we shall assume that the probability distribution of the sources will be modelled by non-Gaussian prior distributions $p_i(s_i)$. Since the relation between sources and inputs is deterministic and one-to-one, we may view it as a change of coordinates. Deriving an expression for the probability distribution of the inputs $\mathbf{x}$ can therefore be accomplished by transforming expression (1) to $\mathbf{x}$-space, using the Jacobian of that transformation,

$$p(\mathbf{x}) = \prod_{i=1}^{M} p_i(s_i(\mathbf{x})) \left| \frac{\partial \mathbf{s}}{\partial \mathbf{x}} \right| = \prod_{i=1}^{M} p_i(\mathbf{w}_i^T \mathbf{x}) |\det W| \tag{3}$$

where $\mathbf{w}_i^T$ are the rows of W. Learning proceeds by averaging the log-likelihood for the above model over a data distribution[2] $p^0(\mathbf{x})$ and using the derivatives of it with respect to W for gradient ascent:

$$\frac{\partial \left\langle \log p(\mathbf{x}) \right\rangle_{p^0}}{\partial w_{ij}} = \left\langle \frac{\partial \log p_i(s_i)}{\partial s_i} x_j \right\rangle_{p^0} + \left[W^{-T} \right]_{ij}$$

where w_{ij} is the j^{th} entry of $\mathbf{w}_i$ and $[W^{-T}]_{ij}$ is the ij^{th} entry of the matrix W^{-T}.

2.2 The Information Maximization Approach

An alternative, more neurally plausible approach to ICA was put forward by Bell and Sejnowski (1995).[3] In that paper it was assumed that a certain transformation was applied to the inputs,

$$y_i = f_i(\mathbf{w}_i^T \mathbf{x}) \qquad i = 1 \ldots M \tag{4}$$

with $f_i(\cdot)$ being a monotone squashing function such as a sigmoid and $\mathbf{w}_i$ a set of linear filters. It was then argued that maximizing the mutual information[4] between outputs $\mathbf{y}$ and inputs $\mathbf{x}$, which is equivalent to maximizing the *entropy* of $\mathbf{y}$ due to the deterministic relation (4), would lead to independent components. This effect can be understood through the decomposition

$$H(\mathbf{y}) = \sum_{i=1}^{M} H_i(y_i) - I(y_1, \ldots, y_M)$$

2. This data distribution is the underlying distribution from which our observed data is sampled. In practice, we replace this by the empirical distribution over the training set.
3. In fact this information maximization approach to ICA was proposed first, followed by the causal generative approach. We presented the two approaches in reverse order here for more intuitive exposition.
4. Note that this mutual information is measured with respect to the data distribution p^0.

with $H(\mathbf{y})$ the entropy of $\mathbf{y}$, $H_i(y_i)$ the individual entropies, and I the mutual information among y_i's. Maximizing the joint entropy thus involves maximizing the individual entropies of the y_i's and *minimizing* the mutual information between the y_i's, i.e. making the y_i's independent.

This approach can best be described as a filtering approach, since each y_i is just a squashed version of the filter outputs $s_i = \mathbf{w}_i^T \mathbf{x}$. This is in contrast with the causal generative approach where we instead think of $\mathbf{x}$ as being generated by $\mathbf{s}$ in a top-down manner.

2.3 Equivalence of the Two Approaches

For square representations the information maximization approach turns out to be *equivalent* to the causal generative one if we interpret $f_i(\cdot)$ to be the cumulative distribution function of $p_i(\cdot)$ (Pearlmutter and Parra, 1996, MacKay, 1996, Cardoso, 1997). This can be seen by observing that the entropy $H(\mathbf{y})$ can be written as a negative KL divergence using a change of variables as follows,

$$H(\mathbf{y}) = -\int d\mathbf{y}\; p^0(\mathbf{y}) \log p^0(\mathbf{y}) = -\int d\mathbf{x}\; p^0(\mathbf{x}) \log \frac{p^0(\mathbf{x})}{|\det J(\mathbf{x})|} \tag{5}$$

where $p^0(\mathbf{y}) = \frac{p^0(\mathbf{x})}{|\det J(\mathbf{x})|}$ and $J(\mathbf{x})$ is the Jacobian of the transformation between $\mathbf{y}$ and $\mathbf{x}$,

$$J_{ij}(\mathbf{x}) = \frac{\partial y_i(\mathbf{x})}{\partial x_j} \tag{6}$$

Using some basic algebra it can be shown that $|\det J(\mathbf{x})|$ is in fact exactly equal to eqn. (3), since f_i satisfies $f_i'(\cdot) = p_i(\cdot)$, being the cumulative function of p_i. Therefore maximizing the entropy in eqn. (5) is indeed equivalent to maximizing the log likelihood of the model (3). Note also that if the sources are *actually* distributed according to $p_i(\cdot)$ and mixed using A, then the transformation (4) maps the input variables to independent, uniformly distributed variables over the range of $f_i(\cdot)$, i.e. the interval $[0, 1]$ in case of a sigmoid. This geometric interpretation will be helpful in Section 3.2.

2.4 Square ICA with Input Noise

In the previous section we have shown that the causal generative approach, in the special case of a square mixing matrix and no noise, is equivalent to the information maximization approach. This equivalence will break down, however, when we consider a noise model for the inputs. In that case, there is no longer a deterministic relationship between the inputs and the sources. It is, however, still straightforward to write a probabilistic model for the joint distribution over sources and inputs,

$$p(\mathbf{x}, \mathbf{s}) = p(\mathbf{x}|\mathbf{s}) \prod_{i=1}^{M} p_i(s_i)$$

Unfortunately, even for isotropic Gaussian noise it is no longer true that given a mixing matrix A the optimal reconstruction of the sources is simply given by eqn. (2). Instead, one typically computes the maximum a posteriori (MAP) value (or the mean, depending on the objective) of the *posterior* distribution $p(\mathbf{s}|\mathbf{x})$.

3. Overcomplete Generalizations of ICA

The equivalence between the two approaches also breaks down when there are more sources than input dimensions, i.e. when we consider overcomplete representations of the data. We will now review overcomplete generalizations of ICA based on both approaches.

3.1 The Causal Generative Approach

Arguably the most natural way to extend the ICA framework to overcomplete representations is through the causal generative approach. The corresponding directed graphical model is depicted in Figure 2A. For noiseless inputs, finding the most probable state $\mathbf{s}$ corresponding to a particular input $\mathbf{x}$ now translates into the following optimization problem:

$$\mathbf{s}^{\mathbf{MAP}} = \underset{\mathbf{s}}{\operatorname{argmax}} \left\{ \sum_{i=1}^{M} \log p_i(s_i) \;\middle|\; \text{such that } \mathbf{x} = A\mathbf{s} \right\}$$

The above problem is typically hard and it can only be solved efficiently for certain choices of $p_i(\cdot)$. For instance Lewicki and Sejnowski (2000) argued that by choosing the priors to be Laplacian the problem can be mapped to a standard linear program.

One can "soften" this optimization problem by introducing a noise model for the inputs. For instance, using a spherical Gaussian noise model with noise variance σ^2 we find the following joint probability density distribution over sources and inputs:

$$p(\mathbf{x}, \mathbf{s}) = p(\mathbf{x}|\mathbf{s})\; p(\mathbf{s}) = \mathcal{N}_{\mathbf{x}}\left[A\mathbf{s}, \sigma^2 I\right] \prod_{i=1}^{M} p_i(s_i). \tag{7}$$

which leads to the following maximization problem to reconstruct the sources from the inputs,

$$\mathbf{s}^{\mathbf{MAP}} = \underset{\mathbf{s}}{\operatorname{argmax}} \left\{ -\frac{1}{2\sigma^2}|\mathbf{x} - A\mathbf{s}|^2 + \sum_{i=1}^{M} \log p_i(s_i) \right\} \tag{8}$$

Maximum likelihood learning for the above noisy model using the EM procedure involves averaging over the posterior distribution $p(\mathbf{s}|\mathbf{x})$. Unfortunately this inference problem is intractable in general and approximations are needed. In the literature one can find a whole range of approximate inference techniques applied to this problem. In Olshausen and Field (1996) the posterior is approximated by a delta function at its MAP value. Thus at every iteration of learning and for every data vector the maximization in eqn. (8) needs to be performed.[5] In Lewicki and Sejnowski (2000) it was argued that the approximation can be significantly improved if a Gaussian distribution around this MAP value was constructed by matching the second derivatives locally (i.e. the Laplace approximation). Attias (1999) and Girolami (2001) use a variational approach which replaces the true posterior with a tractable approximation which is itself adapted to better approximate the posterior. Finally, MCMC sampling methods, such as Gibbs sampling may be employed to solve the inference problem approximately (Olshausen and Millman, 2000).

5. In fact the situation is slightly better using a variational point of view. One can show that one can also improve a bound on the log-likelihood by jointly maximizing over $\mathbf{s}$ and A. We also note that an extra condition on the mixing matrix is needed to prevent it from collapsing to 0.

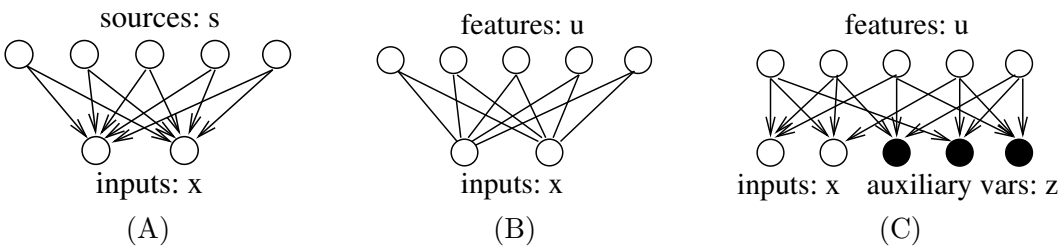

Figure 2: (A) Directed graphical model corresponding to the causal generative approach to ICA. (B) Undirected graphical model for an EBM. (C) Directed graphical model representation for an EBM with auxiliary variables clamped at 0.

A notably different variation on the generative theme is the Bayesian approach taken by Hyvarinen and Inki (2002). There, a prior distribution $p(A)$ over possible mixing matrices A is introduced which favors orthogonal basis vectors (columns of A). They argue that the role of the Jacobian $|\det W| = 1/|\det A|$ in eqn. (3) is precisely to encourage orthogonality among basis vectors and that it is therefore a reasonable assumption to remove this Jacobian in favor of the extra prior. The resultant expression is then easily extended to overcomplete representations.

We want to stress that causal generative models will almost always lead to very difficult inference problems. In contrast, generating unbiased samples from the distribution $p(\mathbf{x})$ is relatively straightforward, since we first sample source values independently from their priors and subsequently sample the input variables according to the conditional Gaussian in eqn. (7).

3.2 The Information Maximization Approach

In Section 2 an information maximization approach to ICA was discussed for the simple case when the number of inputs is equal to the number of sources and no noise is assumed on the inputs. A natural question is whether that objective can be generalized to overcomplete representations. One possibility advocated by Shriki et al. (2002) is to define again the parametrized nonlinear mapping (4) between inputs and outputs and to maximize their mutual information (which amounts to maximizing the entropy of the outputs). Note that this approach is best classified as a filtering approach, and that inputs are mapped one-to-one onto a *subset* of all possible outputs, i.e. the image of that mapping forms a lower dimensional manifold in output space (see Figure 3). Shriki et al. (2002) showed that this objective translates into maximizing the following expression for the entropy,

$$H(\mathbf{y}) = -\int d\mathbf{x}\, p^0(\mathbf{x}) \log \frac{p^0(\mathbf{x})}{\sqrt{\det(J(\mathbf{x})^T J(\mathbf{x}))}} \tag{9}$$

where $J(\mathbf{x})$ is the Jacobian defined in eqn. (6), and $p^0(\mathbf{x})$ is the data distribution.

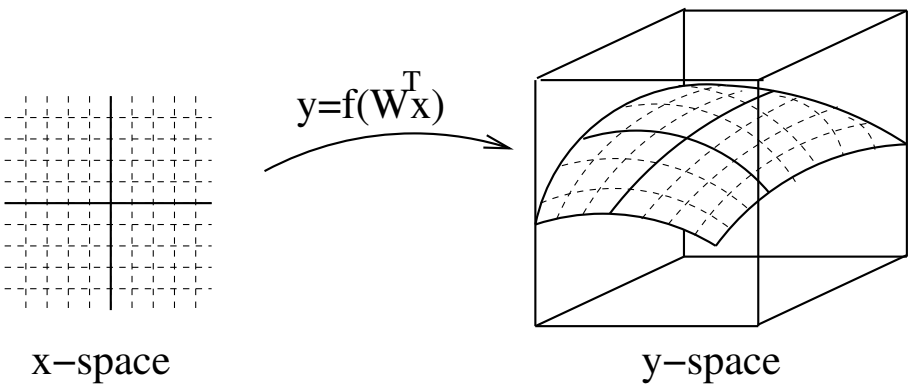

Figure 3: Mapping used by the information maximization approach given by eqn. (4).

4. Energy-Based Models

By interpreting ICA as a filtering model of the inputs, we now describe a very different way of generalizing ICA to overcomplete representations. Energy-based models (EBM) preserve the computationally attractive property that the features $\mathbf{u}$ are simple deterministic functions of the inputs, instead of stochastic latent variables as in a causal generative model. As a consequence, even in the overcomplete setting the posterior $p(\mathbf{u}|\mathbf{x})$ collapses to a point, which stands in sharp contrast to overcomplete causal models which define a posterior *distribution* over the sources. In fact, for overcomplete EBMs, not all feature values are allowed, since not all values lie in the image of the mapping from $\mathbf{x}$ to $\mathbf{u}$. This is similar to the information maximization approach but very different from the causal generative approach where all source values are allowed.

Let $u_i(\mathbf{x}; \mathbf{w}_i)$ be the mapping from $\mathbf{x}$ to feature u_i with parameters $\mathbf{w}_i$. The features are used for assigning an energy $E(\mathbf{x})$, to each possible observation vector $\mathbf{x}$, as follows

$$E(\mathbf{x}) = \sum_{i=1}^{M} E_i(u_i(\mathbf{x}; \mathbf{w}_i)) \tag{10}$$

The probability of $\mathbf{x}$ is defined in terms of its energy through the Boltzmann distribution[6]

$$p(\mathbf{x}) = \frac{e^{-E(\mathbf{x})}}{Z} = \frac{e^{-\sum_i E_i(u_i(\mathbf{x};\mathbf{w}_i))}}{Z} \tag{11}$$

where Z denotes the normalization constant (or partition function),

$$Z = \int_{\mathbf{x}'} e^{-E(\mathbf{x}')} \, d\mathbf{x}' \tag{12}$$

Standard ICA with non-Gaussian priors $p_i(s_i)$ is implemented by having the same number of sources as input dimensions and using

$$u_i(\mathbf{x}, \mathbf{w}_i) = \mathbf{w}_i^T \mathbf{x} \qquad\qquad E_i(u_i) = -\log p_i(u_i) \tag{13}$$

6. We note that the additive form of the energy leads to a product form for the probability distribution, which was called a "product of experts" (PoE) model in (Hinton, 2002).

Furthermore, in this special case of standard ICA the normalization term in eqn. (11) is tractable and simplifies to

$$Z = \left| \frac{1}{\det(W)} \right| \tag{14}$$

where the rows of W are the filters $\mathbf{w}_i^T$.

The above energy-based model suggests thinking about ICA as a filtering model instead of a causal generative model. That is, observations are linearly *filtered* rather than independent sources being linearly *mixed*. Hinton and Teh (2001) interpreted these filters as linear constraints, with the energies serving as costs for violating the constraints. Using energies corresponding to heavy tailed distributions with a sharp peak at zero means that the constraints should be "frequently approximately satisfied", but will not be strongly penalized if they are grossly violated.

In this new approach it is very natural to include more constraints than input dimensions. Note however, that the *marginal independence* among the sources which was a modelling assumption for overcomplete causal models, is no longer true for the features in the EBMs in general. Instead, since the posterior $p(\mathbf{u}|\mathbf{x})$ reduces to a point, the features *given* the inputs are trivially independent,

$$p(\mathbf{u}|\mathbf{x}) = \prod_i \delta(u_i - \hat{u}_i(\mathbf{x}, \mathbf{w}_i))$$

where $\hat{u}_i(\mathbf{x}, \mathbf{w}_i)$ is the feature computed in eqn. (13). The semantics of such probabilistic models is consistent with that of undirected graphical models as depicted in Figure 2B. The above means that inference in EBMs is trivial. On the other hand, sampling from the distribution $p(\mathbf{x})$ is difficult and involves MCMC in general. This is precisely opposite to causal generative models where inference is hard but sampling easy.

4.1 Relating EBMs to Causal Generative ICA

We will now discuss how the proposed overcomplete EBMs relate to the causal generative approach to ICA. Intuitively, an EBM can be interpreted as the conditional distribution obtained from a larger square ICA model when we observe a number of variables. This relationship explains how is it that in EBMs the features $\mathbf{u}$ are conditionally independent given $\mathbf{x}$, but marginally dependent without observations.

In the previous section we have already argued that when the number of input dimensions matches the number of features, an EBM is strictly equivalent to standard ICA as described in Section 2. In the following we will assume that there are more features than input dimensions (i.e. $M > D$).

Consider an ICA model where we have added $M - D$ auxiliary input dimensions $\mathbf{z}$. We will denote the total input space by $\mathbf{v} = [\mathbf{x}, \mathbf{z}]$. We will also add additional filters from the new $\mathbf{z}$ variables to all features and denote them by F, i.e. the total filter matrix is now $G = [W|F]$. We will assume that the new filters are chosen such that G is invertible, i.e. that the new enlarged space is fully spanned. For this enlarged ICA model we can again write the probability distribution as in eqn. (3), here being

$$p(\mathbf{x}, \mathbf{z}) = \prod_{i=1}^{M} p_i(\mathbf{w}_i^T \mathbf{x} + \mathbf{f}_i^T \mathbf{z}) |\det G|$$

where $\mathbf{f}_i^T$ are the rows of F. Next, we write the probability density for the conditional distribution,

$$p(\mathbf{x}|\mathbf{z}) = \frac{p(\mathbf{x},\mathbf{z})}{p(\mathbf{z})} = \frac{p(\mathbf{x},\mathbf{z})}{\int p(\mathbf{x}',\mathbf{z})\,d\mathbf{x}'}$$

where the $|\det G|$ terms have cancelled. If we choose the auxiliary variables $\mathbf{z} = 0$ then this can be written as

$$p(\mathbf{x}|\mathbf{z}=0) = \frac{\prod_i p_i(\mathbf{w}_i^T\mathbf{x})}{\int \prod_i p_i(\mathbf{w}_i^T\mathbf{x}')\,d\mathbf{x}'}$$

The above is of course just an EBM where the partition function is given by,

$$Z = \int \prod_i p_i(\mathbf{w}_i^T\mathbf{x}')\,d\mathbf{x}'$$

Note that the above derivation is independent of the precise choice of the filters F as long as they span the extra dimensions.

In the previous subsection we have seen that an EBM may be interpreted as an undirected graphical model with conditional independence of the features given the inputs. From the above discussion we may conclude that we can also interpret the EBM as a conditional distribution $p(\mathbf{x}|\mathbf{z}=0)$ on a directed graphical model, where $M-D$ auxiliary variables $\mathbf{z}$ have been clamped at 0 (see Figure 2C). By clamping the extra nodes at 0 we introduce dependencies among the features through the phenomenon of "explaining away". In other words, the features are "marginally" dependent when $\mathbf{x}$ is unobserved. When $\mathbf{x}$ is observed, the whole input vector $[\mathbf{x},\mathbf{z}]$ is now observed so the posterior distribution over the features again collapse to a point, trivially implying conditional independence.

4.2 Relating EBMs to Information Maximization

In Section 3.2 we saw that in the information maximization approach to overcomplete representations one maximizes the entropy eqn. (9). The fact that the quantity $\sqrt{\det(J(\mathbf{x})^T J(\mathbf{x}))}$ in that equation is not normalized in general, as opposed to the complete case, prevents expression (9) from being a negative KL divergence. If we therefore define the probability density,

$$p(\mathbf{x}) = \frac{1}{Z}\sqrt{\det(J(\mathbf{x})^T J(\mathbf{x}))} \tag{15}$$

where Z is the normalization constant, then minimizing the KL divergence $KL[p^0||p]$ is equivalent to maximizing the log-likelihood of the model $p(\mathbf{x})$. Importantly, $p(\mathbf{x})$ is consistent with the definition of an EBM if we choose as the energy,

$$E(\mathbf{x}) = -\log\left(\sqrt{\det(J(\mathbf{x})^T J(\mathbf{x}))}\right) = -\frac{1}{2}\operatorname{Tr}\left[\log\left(J(\mathbf{x})^T J(\mathbf{x})\right)\right]$$

The energy-based density model $p(\mathbf{x})$ in eqn. (15) has a simple interpretation in terms of the mapping (4). This mapping is depicted in Figure 3 where it is shown that the $\mathbf{x}$-coordinates define a parametrization of the manifold. It is not hard to show that the distribution $p(\mathbf{x})$ is transformed precisely to a uniform distribution $p(\mathbf{y}) = 1/Z$ on the manifold in $\mathbf{y}$-space, where the normalization constant Z may thus be interpreted as the *volume* of this manifold.

Minimizing the KL divergence $KL[p^0||p]$ can therefore be interpreted as mapping the data to a manifold in a higher dimensional embedding space, in which the data are distributed as uniformly as possible. The relation between information maximization and the above energy-based approach is summarized by the following expression,

$$H(\mathbf{y}) = -KL(p^0(\mathbf{x})||p(\mathbf{x})) + \log\left(\text{Manifold-Volume}\right) \tag{16}$$

The first term describes the "fit" of the model $p(\mathbf{x})$ to data, while the second term is simply the entropy of the uniform distribution $p(\mathbf{y})$ on the manifold. Relative to the energy-based approach, maximizing the mutual information will have a stronger preference to increase the volume of the manifold, since this is directly related to the entropy of $p(\mathbf{y})$. Note that in the square case the manifold is exactly the whole image space $[0,1]^M$, hence its volume is always fixed at 1, and eqn. (16) reduces exactly to the KL divergence $KL(p^0(\mathbf{x})\|p(\mathbf{x}))$.

In the overcomplete case, experiments will have to decide which approach is preferrable and under what circumstances.

5. Parameter Estimation for Energy-Based Models

In Section 4 we proposed energy-based models as probabilistic models for overcomplete representations. We did however not discuss how to fit the free parameters of such models (e.g. filters $\mathbf{w}_i^T$) efficiently to data. In this section we will address that issue.

First, we describe the usual maximum likelihood method of training such models. For overcomplete models, we show that maximum likelihood is not a practical solution, because of the non-trivial partition function. In light of this, we propose another estimation method for energy-based models called contrastive divergence (Hinton, 2002). This is a biased method, but we will show that the bias is acceptably small compared with the gain in efficiency in training overcomplete models, and the ease with which we can generalize the method to new and more intricate models.

Let $p^0(x)$ be the distribution of the observed data, and $p^\infty(x) = p(x)$ be the model distribution given in eqn. (11) (the notation will become apparent later in the section). We would like p^∞ to approximate p^0 as well as possible. The standard measure of the difference between p^0 and p^∞ is the Kullback-Leibler (KL) divergence:

$$KL(p^0\|p^\infty) = \int p^0(x) \log \frac{p^0(x)}{p^\infty(x)}\, dx$$

Because p^0 is fixed, minimizing the KL divergence is equivalent to maximizing the log likelihood of the data under the model p^∞. For energy-based models given by eqn. (11), the derivative of the KL divergence with respect to a weight w_{ij} is

$$\frac{\partial KL(p^0\|p^\infty)}{\partial w_{ij}} = \left\langle \frac{\partial E(x)}{\partial w_{ij}} \right\rangle_{p^0} - \left\langle \frac{\partial E(x)}{\partial w_{ij}} \right\rangle_{p^\infty} \tag{17}$$

where $\langle\cdot\rangle_q$ is the expectation operator under distribution q. Learning can now proceed by using the derivative in eqn. (17) for gradient descent in the KL divergence between the data distribution and the model distribution,

$$\Delta w_{ij} \propto -\frac{\partial KL(p^0\|p^\infty)}{\partial w_{ij}} \tag{18}$$

The above update rule can be understood as lowering the energy surface at locations where there are data (first term in eqn. (17)) and at the same time raising the energy surface at locations where there are no data but the model predicts high probability (second term in eqn. (17)). This will eventually result in an energy surface with low energy (high probability) in regions where there are data present and high energy (low probability) everywhere else.

The second term on the RHS of eqn. (17) is obtained by taking the derivative of the log partition function (eqn. (12)) with respect to w_{ij}. In the square ICA case, the log partition function is exactly given by $\log|\det W^{-1}|$, hence the second term evaluates to the ij^{th} entry of the matrix W^{-T}. However if the model is overcomplete, there is no analytic form for the partition function so exact computation is generally intractable. Instead, since the second term is an expectation under the model distribution p^∞, one possibility is to use Markov chain Monte Carlo (MCMC) techniques to approximate the average using samples from p^∞ (see Neal, 1993). This method inherits both the advantages and drawbacks associated with MCMC sampling. The obtained estimate is consistent (i.e. the bias decreases to zero as the length of the chains is increased), and it is very easily adaptable to other more complex models. The main drawback is that the method is very expensive – the Markov chain has to be run for many steps before it approaches the equilibrium distribution p^∞ and it is hard to estimate how many steps are required. Also, the variance of the MCMC estimator is usually high. To reduce the variance many independent samples are needed, incurring additional computational costs. Therefore, estimating the derivative (17) accurately by MCMC sampling is slow and can be unreliable due to high variance.

However, in the following we will argue that it is unnecessary to estimate the derivatives averaged over the equilibrium distribution in order to train an energy-based model from data. Instead, we will average the derivatives over a different distribution, resulting from truncating the Markov chain after a fixed number of steps. This idea, called contrastive divergence learning, was first proposed by Hinton (2002) to improve both computational efficiency and reduce the variance at the expense of introducing a bias for the estimates of the parameters with respect to the maximum likelihood solution.

There are two ideas involved in contrastive divergence learning. The first one is to start the Markov chain at the data distribution p^0 rather than to initialize the Markov chain at some vague distribution (e.g. a Gaussian with large variances). The reason usually given for using vague initial distributions is that every mode of the equilibrium distribution has a chance of being visited by some chain. This can help to overcome a problematic feature of many Markov chains – a low mixing rate; once a chain enters a mode of the distribution it is hard to escape to a different mode. However, we argue that starting at the data distribution is preferable since the training data already contains examples from the various modes that the model distribution ought to have. Towards the end of learning, when the modes of the model distribution roughly correspond to the modes in the data distribution, the number of samples in each mode approximately matches the number of data vectors in each mode. This further reduces the variance of the derivative estimates. A possible danger with this technique is that certain spurious (empty) modes which are accidentally created during learning may go unnoticed.

The second idea of contrastive divergence is to run the Markov chain for only a few iterations rather than until equilibrium. Because the chains are started at the data distribution, even after only a few iterations, any consistent tendency to move away from the

data distribution provides valuable information that can be used to adapt the parameters of the model. Intuitively, the parameters of the model should be updated so that the Markov chain does not tend to move away from the data distribution (since we want the Markov chain to equilibrate to the data distribution).

Combining the two ideas described above and defining $p^n(x)$ to be the distribution of the random variable at the n^{th} iteration of the Markov chain,[7] the contrastive divergence learning algorithm is implemented by using the following quantity to update the filters w_{ij},

$$\Delta w_{ij} \propto -\left\langle \frac{\partial E(x)}{\partial w_{ij}} \right\rangle_{p^0} + \left\langle \frac{\partial E(x)}{\partial w_{ij}} \right\rangle_{p^n} \tag{19}$$

Relative to maximum likelihood learning (eqns. (17) and (18)) we have replaced the equilibrium distribution p^∞ with p^n, and the Markov chain is initialized at the data distribution p^0. The figure below gives pseudo-code for contrastive divergence learning.

Contrastive Divergence Learning for Energy-Based Models

1. Compute the gradient of the total energy with respect to the parameters and average over the data cases $\mathbf{d}_k$.
2. Run MCMC samplers for n steps, starting at every data vector $\mathbf{d}_k$, keeping only the last sample $\mathbf{s}_k$ of each chain.
3. Compute the gradient of the total energy with respect to the parameters and average over the samples $\mathbf{s}_k$.
4. Update the parameters using,

$$\Delta w_{ij} = -\frac{\eta}{N}\left(\sum_{\text{data } \mathbf{d}_k} \frac{\partial E(\mathbf{d}_k)}{\partial w_{ij}} - \sum_{\text{samples } \mathbf{s}_k} \frac{\partial E(\mathbf{s}_k)}{\partial w_{ij}}\right) \tag{20}$$

where η is the learning rate and N the number of samples in each mini-batch.

Notice that in order to compute the average in the second term of eqn. (19) we used samples produced by Markov chains initialized at the corresponding data vectors used in the first term. This, rather than uniformly sampling the initial states of the Markov chains from the data vectors, further reduces the variance.

If in addition to the filter weights w_{ij} additional parameters are present, for instance to model the shape of the energies E_i, similar update rules as eqn. (20) can be used to fit them to data. For standard ICA, this would correspond to learning the shape of the prior densities.

In the ideal situation that the model distribution p^∞ is flexible enough to perfectly model the data distribution[8] p^0, and we use a Markov chain that properly mixes, then

7. Hence the notation p^0 for the initial distribution of the Markov chain and p^∞ for the limit distribution of p^n as $n \to \infty$.
8. In the case of finite data, we replace the data distribution by the *empirical* distribution, which is a mixture of delta-functions. In this case, any smooth model distribution will not be able to perfectly fit the empirical data distribution and the above argument fails. In fact, we may expect to incur a certain bias with respect to the maximum likelihood solution.

contrastive divergence learning has a fixed point at the maximum likelihood solution, i.e. when $p^\infty = p^0$. This is not hard to see, since at the maximum likelihood solution, the Markov chain will not change the model distribution, which implies that the derivatives in eqn. (19) precisely cancel. In general however, we expect contrastive divergence learning to trade-off variance with bias (see also Williams and Agakov, 2002). Apart from this, it may also happen that for certain Markov chains spurious fixed points exist in contrastive divergence learning (for some examples see MacKay, 2001).

Although we have argued that contrastive divergence learning seems a sensible way to fit energy-based models to data, we have not shown that it corresponds to gradient descent on a cost function, which is desirable to prove convergence. Now we will show a slightly weaker statement, namely that the update (19) corresponds to an approximate gradient descent step on a cost function. Define the contrastive divergence cost function (Hinton, 2002) as,

$$CD = KL(p^0 \| p^\infty) - KL(p^n \| p^\infty)$$

Note that this consists of the usual KL divergence between the data distribution and the model distribution, subtracted by the KL divergence between the n-step distribution p^n and the model distribution. Using properties of Markov chains one can show that the n-step distribution is always closer to the equilibrium model distribution, so that CD is always non-negative, with $CD = 0$ exactly when $p^0 = p^\infty$.

Taking derivatives of the contrastive divergence cost function with respect to the filter weights w_{ij} we find the following gradient,

$$\frac{\partial CD}{\partial w_{ij}} = \left\langle \frac{\partial E(x)}{\partial w_{ij}} \right\rangle_{p^0} - \left\langle \frac{\partial E(x)}{\partial w_{ij}} \right\rangle_{p^n} - \frac{\partial KL(p^n \| p^\infty)}{\partial p^n} \frac{\partial p^n}{\partial w_{ij}} \tag{21}$$

The first two terms in eqn. (21) are identical to the ones proposed for the learning algorithm in eqn. (19). The last term represents the effect that changes in w_{ij} have on the contrastive divergence via the effect on p^n, i.e the effect on the Markov chain itself when the parameters w_{ij} are altered. This term is hard to compute but fortunately it is typically very small and simulations by Hinton (2002) suggest that it can be safely ignored. The results later in this paper further support this claim.

6. Experiment: Blind Source Separation

To assess the performance of contrastive divergence as a learning algorithm, we compared a hybrid Monte Carlo implementation of contrastive divergence with an exact sampling algorithm as well as the Bell and Sejnowski (1995) algorithm on a standard "blind source separation" problem. The model has the same number of input and source dimensions,[9] and the energy of the model is defined as

$$E_i(s_i) = -\log(\sigma(s_i)(1 - \sigma(s_i)))$$

where $\sigma(s) = 1/(1 + \exp(-s))$ is the sigmoid function. This model is strictly equivalent to the noiseless ICA model with sigmoidal outputs used by Bell and Sejnowski (1995).

9. Note however that recovering more sound sources than input dimensions (sensors) is not possible with our energy-based model, since the feautures are not marginally independent.

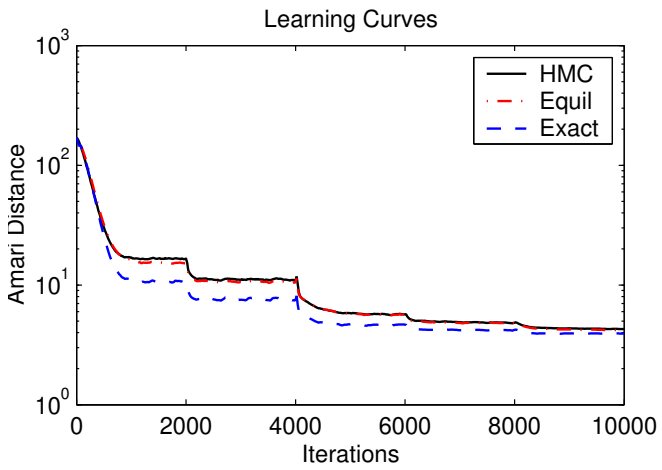

Figure 4: Evolution of the Amari distance for the various algorithms, averaged over 100 runs. Note that HMC converged just as fast as the exact sampling algorithm Equil, while the exact algorithm Exact is only slightly faster. The sudden changes in Amari distance are due to the annealing schedule.

The data consisted of 16, 5-second stereo CD recordings of music, sampled at 44.1 kHz.[10] Each recording was monoized, down-sampled by a factor of 5, randomly permuted over the time-index and rescaled to unit variance. The resulting 88436 samples in 16 channels were linearly mixed using the standard *instamix* routine with $b = 0.5$ (1 on the diagonal and 1/9 off the diagonal),[11] and whitened before presentation to the various learning algorithms.

We compared three different ways of computing or estimating the gradient (17):

Algorithm HMC: We used a hybrid Monte Carlo implementation of contrastive divergence. This implementation uses 1 step of hybrid Monte Carlo simulation to sample from $p^1(\mathbf{x})$, which in turn consists of 30 leap frog steps, with the step sizes adapted at the end of each simulation so that the acceptance rate is about 90%. See Neal (1993) for further detail on hybrid Monte Carlo.

Algorithm Equil: For noiseless ICA, it is possible to sample efficiently from the true equilibrium distribution using the causal generative view. These samples can then be used to estimate the second term of (17). To be fair, we used a number of samples equal to the number of data vectors in each mini-batch.

Algorithm Exact: We can also compute the partition function using eqn. (14), and evaluate the second term of eqn. (17) exactly. This is precisely Bell and Sejnowski's algorithm.

Parameter updates were performed on mini-batches of 100 data vectors. The learning rate was annealed from 0.05 down to 0.0005 in 10000 iterations of learning,[12] while a momentum

10. Prepared by Barak Pearlmutter and available at `http://sweat.cs.unm.edu/~bap/demos.html`.
11. Available at `http://sound.media.mit.edu/ica-bench/`.
12. 2000 iterations each at 0.05, 0.025, 0.005, 0.0025 and 0.0005.

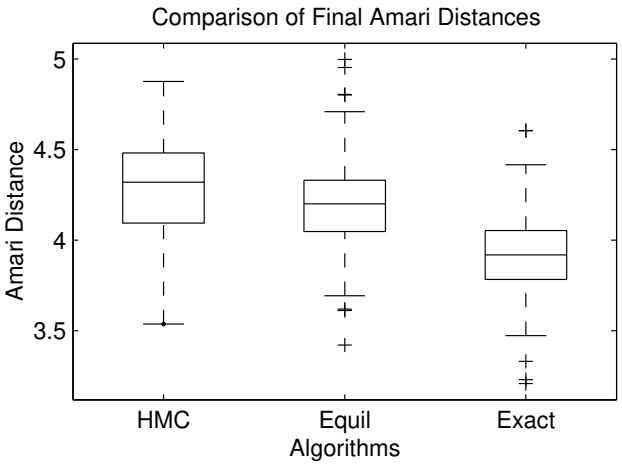

Figure 5: Final Amari distances for the various algorithms, averaged over 100 runs. The boxes have lines at the lower quartile, median, and upper quartile values. The whiskers show the extent of the rest of the data. Outliers are denoted by "+". This plot shows that the deterministic method Exact performs slightly better than the sampling methods HMC and Equil, probably due to the variance induced by the sampling. More importantly, it shows that learning with brief sampling (HMC) performs equally well as learning with samples from the equilibrium distribution (Equil).

factor of 0.9 was used to speed up convergence. The initial weights were sampled from a Gaussian with a standard deviation of 0.1.

During learning we monitored the Amari distance[13] to the true unmixing matrix. In Figures 4 and 5 we show the results of the various algorithms on the sound separation task. The main conclusion of this experiment is that we do not need to sample from the equilibrium distribution in order to learn the filters $\mathbf{W}$. This validates the ideas behind CD learning.

7. Experiments: Feature Extraction

We present examples of the features delivered by our algorithm on several standard datasets. Firstly we demonstrate performance on typical ICA tasks of determining an overcomplete set of features of speech and natural images. Then, we show the algorithm applied to the CEDAR cdrom dataset of handwritten digits and lastly, we present the feature vectors learned when the algorithm is applied to the FERET database of human faces.

For all the experiments described in this section we use an energy function of the form:

$$E_i(u_i(\mathbf{x}, \mathbf{w}_i)) = \gamma_i \log\left(1 + (\mathbf{w}_i^T \mathbf{x})^2\right)$$

13. The Amari distance (Amari et al., 1996) measures a distance between two matrices A and B up to permutations and scalings: $\left(\sum_{i=1}^{N}\sum_{j=1}^{N} \frac{|(AB^{-1})_{ij}|}{\max_k |(AB^{-1})_{ik}|} + \frac{|(AB^{-1})_{ij}|}{\max_k |(AB^{-1})_{kj}|}\right) - 2N^2$.

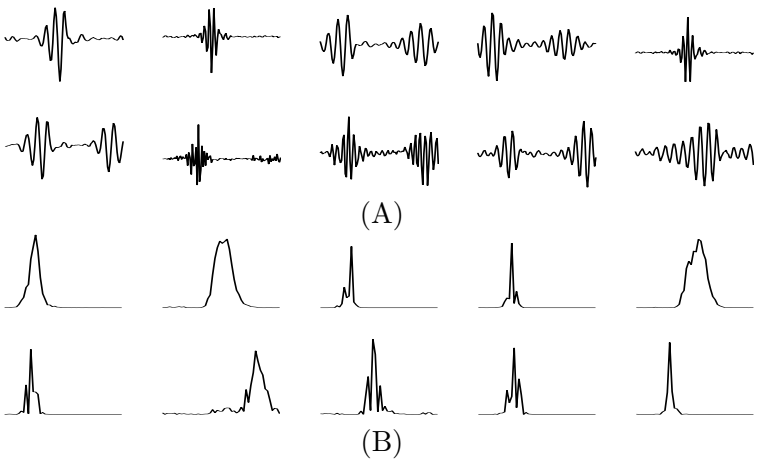

Figure 6: (A) Filters found by the 2× overcomplete EBM. The 5 filters in the first row are the ones with largest power, indicating that they represent important features. The 5 filters in the second row are randomly drawn from the remaining 195 filters. (B) Corresponding power-spectra.

which corresponds to modelling the data with a product of one-dimensional student-t distributions of degree $(2\gamma - 1)$ (Hinton and Teh, 2001). This energy function was chosen for its simplicity yet versatility in describing super-Gaussian distributions. However, the algorithmic formulation allows the use of arbitrary energy functions and results may be improved by a more systematic tailoring of the energy function to particular datasets.

7.1 Speech

To test whether the model could extract meaningful filters from speech data we used recordings of 10 male speakers from the TIMIT database, uttering the sentence:

"Don't ask me to carry an oily rag like that."

The sentences were down-sampled to 8kHz and 50000, 12.5ms segments (each segment corresponding to 100 samples) were extracted from random locations. Before presentation to the learning algorithm the data was centred and sphered. The features were trained using contrastive divergence with one step of hybrid Monte Carlo sampling consisting of 20 leap frog steps. Mini-batches of size 100 were used, while the learning rate was annealed from 0.05 to 0.0005 over 20000 iterations. The filters were initialized at small random values and momentum was used to speed up convergence.

In Figure 6 we show 10 of the 200 features in the whitened domain together with their power spectra. Recall that since there are 2 times more filters extracted as dimensions in the input space, the energy-based model is no longer equivalent to a causal ICA model. Figure 7 shows the distribution of power over time and frequency. There seems to be interesting structure around 1.5kHz, where the filters are less localized and more finely tuned in frequency than average. This phenomenon is also reported by Abdallah and Plumbley (2001).

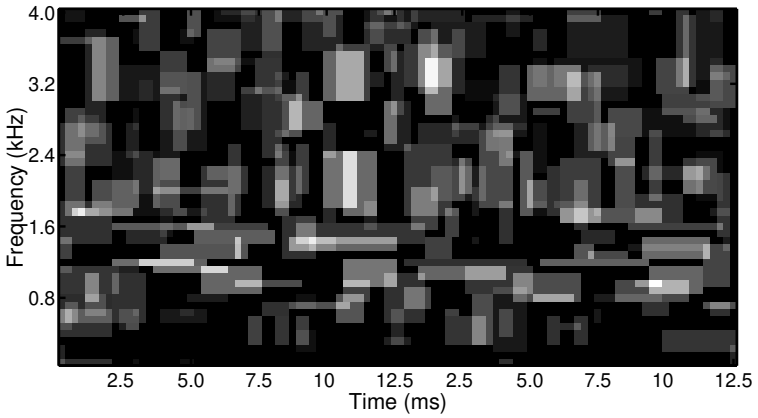

Figure 7: Distribution of power over time and frequency. First the envelope of each filter (the absolute value of its Hilbert transform) was computed and squared. Next, the squared envelope and the power spectrum were thresholded by mapping all values greater than half the peak value to one and the rest to zero. Gaps smaller than 6 samples in time and 3 samples in frequency were filled in. Finally, the outer product of the two "templates" were computed, weighted by the total power of the filter, and added to the diagram.

7.2 Natural Image Patches

We tested our algorithm on the standard ICA task of determining the "independent" components of natural images. The data set used is the `imlog`[14] data set of van Hateren and van der Schaaf (1998). The logarithm of the pixel intensities was first taken and then the image patches were centred and whitened. There were 122880 patches and each patch was 16×16 in size. We trained a network with $256 \times 3 = 768$ features, using contrastive divergence with 1 step of hybrid Monte Carlo sampling consisting of 30 leap frog steps. The step size was adaptive so that the acceptance rate is approximately 90%. Both $\mathbf{w}_i$ and γ_i are unconstrained, but a small weight decay of 10^{-4} was used for $\mathbf{w}_i$ to encourage the features to localize. The $\mathbf{w}_i$'s were initialized to random vectors of length 1, while the γ_i's were initialized at 1. Both $\mathbf{w}_i$ and γ_i were trained with a learning rate of 0.01 and momentum factor of 0.9. We found however that the result is not sensitive to the settings of these parameters. A random sample of 144 learned features in the whitened domain is shown in Figure 8. They were roughly ordered by increasing spatial frequency. By hand, we counted a total of 19 features which have not localized either in the spatial or frequency domain. Most of the other features can be described well with Gabor functions. To further analyze the set of learned filters, we fitted a Gabor function of the form used by Lewicki and Olshausen (1999) to each feature and extracted parameters like frequency, location and extent in the spatial and frequency domains. These are summarized in Figures 9 and 10, and show that the filters form a nice tiling of both the spatial and frequency domains. We see from Figures 9 and 10 that filters are learned at multiple scales, with larger features

14. Available at `ftp://hlab.phys.rug.nl/pub/samples/imlog`.

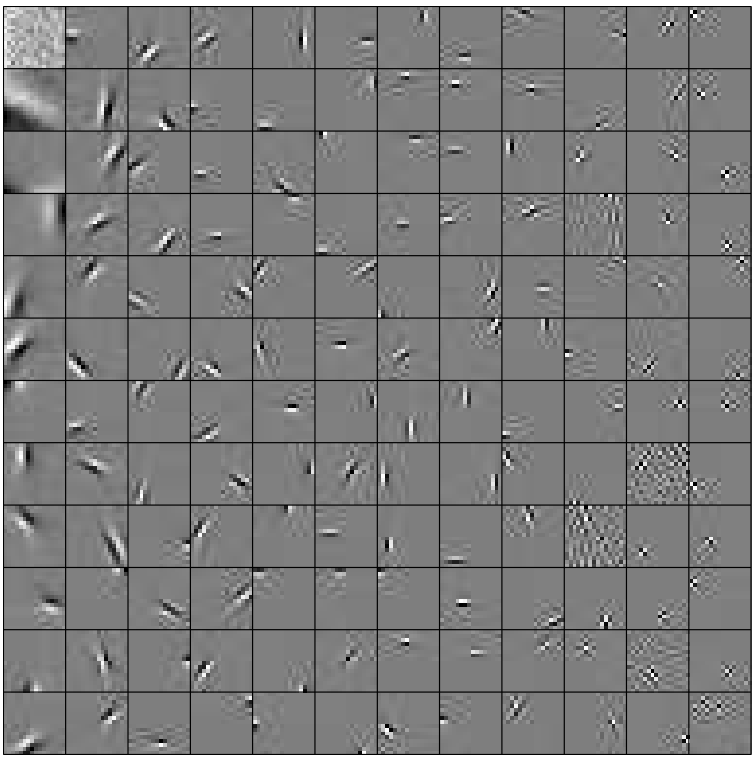

Figure 8: Learned filters for natural images

typically being of lower frequency. However we also see an over emphasis of horizontal and vertical filters. This effect has been observed in previous papers (van Hateren and van der Schaaf, 1998, Lewicki and Olshausen, 1999), and is probably due to pixellation.

7.3 CEDAR Digits

We used 16x16 real valued digits from the "br" set on the CEDAR cdrom #1. There are 11000 digits available, divided equally into 10 classes. The mean image from the entire dataset was subtracted from each datum, and the digits were whitened with ZCA. A network with 361 features was trained in the same manner as for natural image patches.

A random subset of learned filters is shown in Figure 11. To make it easier to discern the structure of the learned filters, we present them in the ZCA whitened domain rather than in pixel space. We note the superficial similarity between these filters and those found from the natural scene experiments. However, in addition to straight edge filters we also see several curved filters. We interpret the results as a set of 'stroke' detectors, modelling a space of strokes that gives rise to the full digit set.

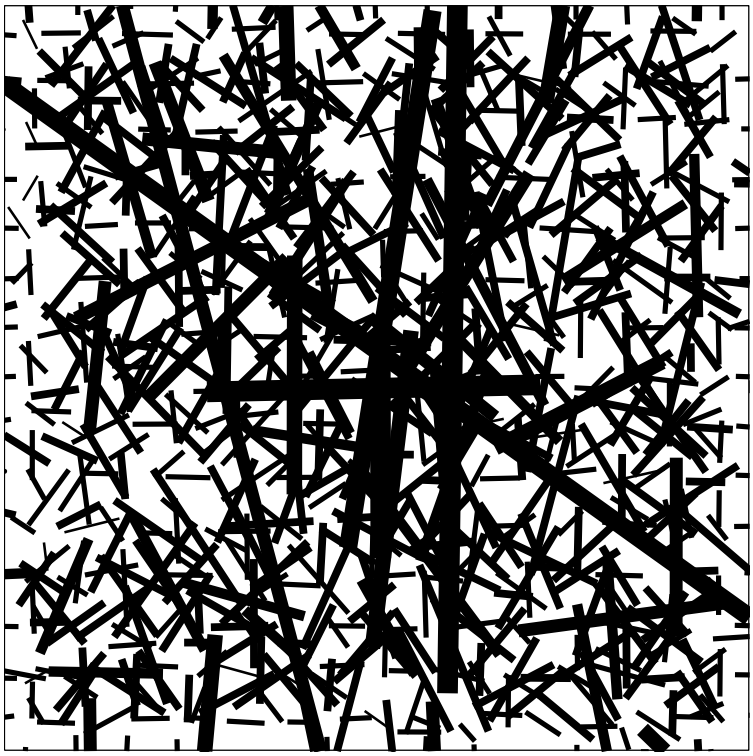

Figure 9: The spatial layout and size of the filters, which are described by the position and size of the bars.

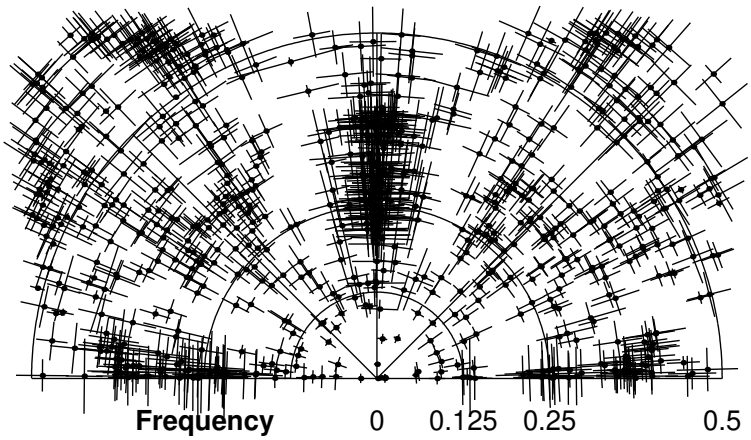

Figure 10: A polar plot of frequency tuning and orientation selectivity of the learned filters, with the centre of each cross at the peak frequency and orientation response, and crosshairs describing the 1/16-bandwidth.

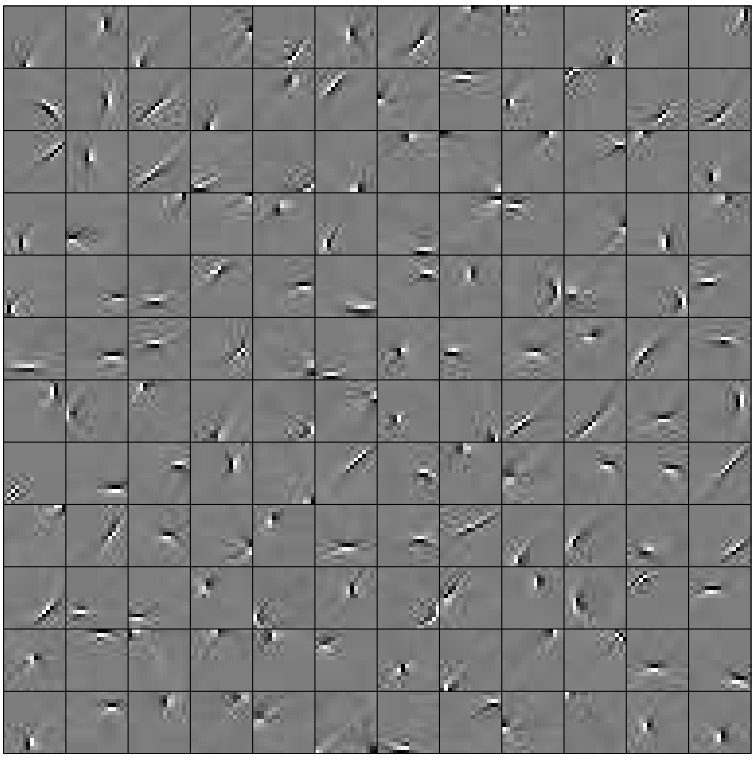

Figure 11: Learned filters for CEDAR digits. Filters are plotted in whitened space for clarity.

7.4 FERET Faces

We used the full NIST FERET database of frontal face images.[15] The data was first preprocessed in the standard manner by aligning the faces, normalising the pixel intensities and cropping a central oval shaped region.[16] Then as an additional preprocessing step we centred the data and performed PCA whitening, retaining the projections onto the leading 256 eigenvectors as the input dimensions to the algorithm. We trained a network with 361 features using contrastive divergence with hybrid Monte Carlo sampling (3 sets of 20 leap frog steps). Both the $\mathbf{w}_i$'s and γ_i's were unconstrained. The $\mathbf{w}_i$'s were initialized uniformly over vectors of norm 1. The γ_i's were initialized at 1. A learning rate of 0.005 was used for the $\mathbf{w}_i$, whilst a learning rate of 0.001 was used for the γ_i.

Figure 12A shows the 32 leading eigenvectors plotted as face images, and Figure 12B shows a subset of 32 filters as learned by our algorithm. In Bartlett et al. (2002) two kinds of square ICA were applied to a lower resolution version of the FERET database. In relation to their work, the filters shown in Figure 12B correspond to 'type II' ICA (each face constitutes an input line) rather than 'type I' ICA (the value of a pixel across all faces constitutes an

15. `http://www.itl.nist.gov/iad/humanid/feret`.
16. `http://www.cs.coloradostate.edu/evalfacerec/index.htm`.

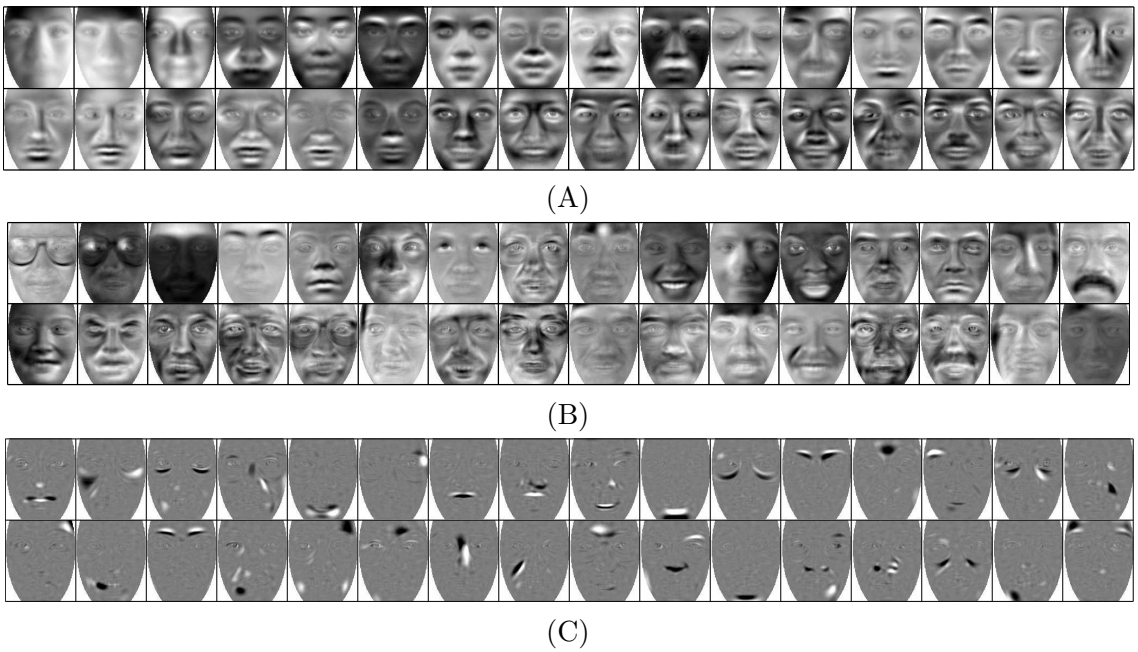

Figure 12: (A) 32 eigenfaces with largest eigenvalue plotted rowwise in descending eigenvalue order. (B) Subset of 32 'type II' feature vectors. The top row are hand picked, the bottom row randomly selected. (C) Subset of 32 'type I' feature vectors, all randomly selected

input line.) There seem to be both similarities and notable differences between our features and their 'type II' results. As with Bartlett et al. (2002), many of the filters that we learn are somewhat global in the sense that most pixels having a nonzero weight. However, in addition to these global features and in contradistinction to their 'type II' results, we also develop features with most of their weight concentrated in localised sub-regions – for instance focusing on glasses, eyes, smiling mouths, moustaches, etc. Furthermore, as well as global features that can perhaps be described as 'archetypical faces' we also see global features which appear to mainly capture structure in the illumination of a face.

Lastly, Figure 12C illustrates the results when our algorithm is applied in a 'type I' manner (using the pixels of the leading 256 principal components as inputs rather than the original pixel values). This approach leads to features that are all highly localised in space. Our results are again qualitatively similar to those described by Bartlett et al. (2002).

Future work will show whether the overcomplete feature sets that our algorithm delivers can be usefully employed in a face or expression recognition system.

8. Discussion

In this paper we have re-interpreted the standard ICA algorithm as an energy-based model and studied its extension to overcomplete representations. We have shown that parameters of an EBM, such as the filter weights and those parametrizing the energy function,

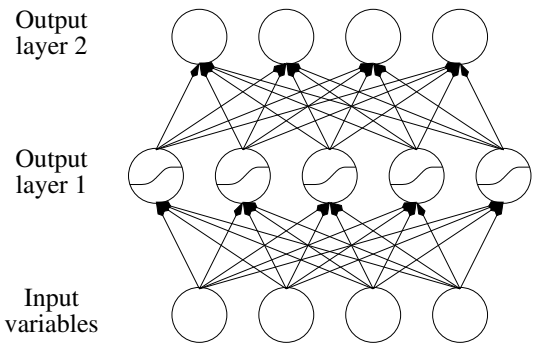

Figure 13: Architecture of a hierarchical non-linear energy-based model. Non-linearities are indicated by sigmoidal units in output layer 1. Energies can be contributed by output variables in both layers, and the number of output variables need not correspond to the number of input variables.

can be efficiently estimated using contrastive divergence learning. Through a number of experiments on standard data sets we have shown that EBMs can efficiently extract useful features in high dimensions.

Contrary to causal generative models for overcomplete ICA, the features of an EBM exhibit marginal dependencies. The advantage of allowing these dependencies in the model is fast inference. In causal generative models, the assumption of marginal independence often leads to intractible inference which needs to be approximated using some iterative, data dependent scheme. The role of these iterations can be understood as suppressing the "activity" of less relevant features, thus producing a sparse code. Therefore, for causal generative models, overcomplete representations are expected to produce very compact (or sparse) codes, a fact which is often emphasized as desirable (Olshausen and Field, 1997). Perhaps surprisingly, we have shown that such a slow iterative process is not required for producing sparse and overcomplete representations.

However the above does suggest enriching EBMs with inhibitory lateral connections to achieve the goal of further suppressing less relevant features in order to produce an even sparser representation. Preliminary experiments using a mean field approach to implement these lateral inhibitions have been successful in learning good density models, but are slow due to the iterative optimization for every data case.

Another powerful generalization of EBMs is a hierarchical non-linear architecture in which the output activities are computed with a feed-forward neural network (Figure 13) and each layer may contribute to the total energy (for related work see also Hyvrinen and Hoyer, 2001). To fit this model to data, backpropagation is used to compute gradients of the energy with respect to both the data vector (to be used in Hybrid Monte Carlo sampling), and the weights (to be used for weight updates). Since this algorithm applies backpropagation in an unsupervised setting and combines it with contrastive divergence learning we have named it "contrastive backpropagation" (Hinton et al., 2003).

Indeed, the contrastive backpropagation learning procedure is quite flexible. It puts no constraints other than smoothness on the activation functions or the energy functions.[17] The procedure can be easily modified to use recurrent neural networks that contain directed cycles by running each forward pass for some predetermined number of steps and defining the energy to be any smooth function of the time history of the activations. Backpropagation through time (Rumelhart et al., 1986, Werbos, 1990) can then be used to obtain the required derivatives. The data-vector can also change during the forward pass through a recurrent network. This makes it possible to model sequential data, such as video sequences, by running the network forward in time for a whole sequence and then running it backward in time to compute the derivatives required for Hybrid Monte Carlo sampling and for updating the weights.

In Welling et al. (2003), a two-layer model was studied where the second layer performed a local averaging of the non-linearly transformed activities of the first layer. This resulted in a topographic ordering of the filters, where orientation, location and frequency are changing smoothly from one filter to the next.

The energy-based approach to ICA we have presented stems from previous work on products of experts (PoEs) (Hinton, 2002). In fact our model is a type of PoE in which each energy term corresponds to one expert.

There is also an interesting link between EBMs and maximum entropy models (Della Pietra et al., 1997, Zhu et al., 1997). Indeed, the probability distribution of a maximum entropy model is also defined as a Boltzmann distribution eqn. (11) over a sum of energy contributions $E_i(\mathbf{x})$ which are written as,

$$E_i(\mathbf{x}) = \lambda_i u_i(\mathbf{x})$$

where $u_i(\mathbf{x})$ are fixed features of the model and where the weights λ_i are the parameters of the model to be fit to data. In this sense, the proposed EBM can be interpreted as a maximum entropy model with flexible learned features and a different energy function.

In conclusion we believe that the EBM provides a flexible modelling tool which can be trained efficiently to uncover useful structure in data.

Acknowledgments

We would like to thank Peter Dayan, Sam Roweis, Zoubin Ghahramani and Maneesh Sahani for helpful discussions, Carl Rasmussen for making *minimize.m* available, and the reviewers and Dave MacKay for helpful comments.

References

S. A. Abdallah and M. D. Plumbley. If edges are the independent components of natural images, what are the independent components of natural sounds? In *International Conference On Independent Component Analysis and Blind Signal Separation*, 2001.

17. There is also the necessary assumption that the energy is defined such that the Boltzmann distribution is normalizable.

S. Amari, A. Cichocki, and H. Yang. A new algorithm for blind signal separation. In *Advances in Neural Information Processing Systems*, volume 8, pages 757–763, 1996.

H. Attias. Independent Factor Analysis. *Neural Computation*, 11:803–851, 1999.

M. S. Bartlett, J. R. Movellan, and T. J. Sejnowski. Face recognition by independent component analysis. *IEEE Transactions on Neural Networks*, 2002. In Press.

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

J. F. Cardoso. Infomax and maximum likelihood for blind source separation. *IEEE Signal Processing Letters*, 4(4):112–114, 1997.

S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998.

P. Comon. Independent component analysis: A new concept? *Signal Processing*, 36: 287–314, 1994.

S. Della Pietra, V. Della Pietra, and J. Lafferty. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(4):380–393, 1997.

M. Girolami. A variational method for learning overcomplete representations. *Neural Computation*, 13:2517–2532, 2001.

G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

G. E. Hinton and Y. W. Teh. Discovering multiple constraints that are frequently approximately satisfied. In *Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence*, pages 227–234, 2001.

G. E. Hinton, Y. W. Teh, M. Welling, and S. Osindero. Contrastive backpropagation. In preparation, 2003.

A. Hyvarinen and M. Inki. Estimating overcomplete independent component bases for image windows. *Journal of Mathematical Imaging and Vision*, 2002. in press.

A. Hyvrinen and P. O. Hoyer. A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images. *Vision Research*, 41(18):2413–2423, 2001.

M. S. Lewicki and B. A. Olshausen. A probabilistic framework for the adaptation and comparison of image codes. *Journal of the Optical Society of America A: Optics, Image Science, and Vision*, 16(7):1587–1601, 1999.

M. S. Lewicki and T. J. Sejnowski. Learning overcomplete representations. *Neural Computation*, 12:337–365, 2000.

D. MacKay. Maximum likelihood and covariant algorithms for independent components analysis. Available electronically at `http://www.inference.phy.cam.ac.uk/mackay/abstracts/ica.html`, 1996.

D. MacKay. Failures of the one-step learning algorithm. Available electronically at `http://www.inference.phy.cam.ac.uk/mackay/abstracts/gbm.html`, 2001.

S. G. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions in Signal Processing*, 41(12):3397–3414, 1993.

R. M. Neal. Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG-TR-93-1, University of Toronto, Department of Computer Science, 1993.

B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1. *Vision Research*, 37:3311–3325, 1997.

B. A. Olshausen and K. J. Millman. Learning sparse codes with a mixture-of-gaussians prior. In *Advances in Neural Information Processing Systems*, volume 12, pages 841–847, 2000.

B. A. Pearlmutter and L. C. Parra. A context-sensitive generalization of ICA. In *Proceedings of the International Conference on Neural Information Processing*, 1996.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, and the PDP Research Group, editors, *Parallel Distributed Processing: Explorations in The Microstructure of Cognition. Volume 1: Foundations*. The MIT Press, 1986.

O. Shriki, H. Sompolinsky, and D. D. Lee. An information maximization approach to overcomplete and recurrent representations. In *Advances in Neural Information Processing Systems*, pages 612–618, 2002.

E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger. Shiftable multi-scale transforms. *IEEE Transactions Information Theory*, 38(2):587–607, 1992.

J. H. van Hateren and A. van der Schaaf. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings of the Royal Society of London B*, 265:359–366, 1998.

M. Welling, G. E. Hinton, and S. Osindero. Learning sparse topographic representations with products of student-t distributions. In *Advances in Neural Information Processing Systems*, 2003.

P. J. Werbos. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560, 1990.

C. K. I. Williams and F. V. Agakov. An analysis of contrastive divergence learning in gaussian boltzmann machines. Technical Report EDI-INF-RR-0120, Institute for Adaptive and Neural Computation, University of Edinburgh, 2002.

S. C. Zhu, Y. N. Wu, and D. Mumford. Minimax entropy principle and its application to texture modelling. *Neural Computation*, 9(8), 1997.