Learning the Structure of Deep Sparse Graphical Models

Ryan Prescott Adams
University of Toronto

Hanna M. Wallach
University of Massachusetts Amherst

Zoubin Ghahramani
University of Cambridge

Abstract

Deep belief networks are a powerful way to model complex probability distributions. However, it is difficult to learn the structure of a belief network, particularly one with hidden units. The Indian buffet process has been used as a nonparametric Bayesian prior on the structure of a directed belief network with a single infinitely wide hidden layer. Here, we introduce the cascading Indian buffet process (CIBP), which provides a prior on the structure of a layered, directed belief network that is unbounded in both depth and width, yet allows tractable inference. We use the CIBP prior with the nonlinear Gaussian belief network framework to allow each unit to vary its behavior between discrete and continuous representations. We use Markov chain Monte Carlo for inference in this model and explore the structures learned on image data.

1 Introduction

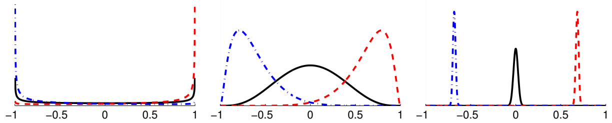
The belief network or directed probabilistic graphical model (Pearl, 1988) is a popular and useful way to represent complex probability distributions. Methods for learning the parameters of such networks are well-established. Learning network structure, however, is more difficult, particularly when the network includes unobserved hidden units. Then, not only must the structure (edges) be determined, but the number of hidden units must also be inferred. This paper contributes a novel nonparametric Bayesian perspective on the general problem of learning graphical models with hidden variables. Nonparametric Bayesian approaches to this problem are appealing because they can avoid the difficult computations required for selecting the appropriate *a posteriori* dimensionality of the model. Instead, they introduce an infinite number of parameters into the model *a priori* and the inference procedure determines the subset of these parameters that actually contributed to the observations. The Indian buffet process (IBP) (Griffiths and Ghahramani, 2006) is one example of a nonparametric Bayesian prior. It has previously been used to introduce an infinite number of hidden units into a belief network with a single hidden layer (Wood et al., 2006) or with a pre-specified number of layers (Courville et al., 2009).

This paper unites two important areas of research: nonparametric Bayesian methods and deep belief networks. Specifically, we develop a nonparametric Bayesian framework to perform structure learning in deep networks, a problem that has not been addressed to date. We first propose a novel extension to the Indian buffet process—the cascading Indian buffet process (CIBP)—and use the Foster-Lyapunov criterion to prove convergence properties that make it tractable with finite computation. We then use the CIBP to generalize the single-layered, IBP-based, directed belief network to construct multi-layered networks that are both infinitely wide and infinitely deep. We discuss useful properties of such networks including expected in-degree and out-degree for individual units. Finally, we combine this approach with the continuous sigmoidal belief network framework of Frey (1997). This framework allows us to infer the type (i.e., discrete or continuous) of individual hidden units—an important property that is not widely discussed in previous work. In summary, we present a flexible, nonparametric framework for directed deep belief networks that permits inference of the number of hidden units, the directed edge structure between units, the depth of the network, and the most appropriate type for each unit.

2 Finite Belief Networks

We consider belief networks that are layered directed acyclic graphs with both visible and hidden units. Hidden units are random variables that appear in the joint distribution described by the belief network but are not observed. We index layers by m, increasing with depth up to M, and allow visible units (i.e., observed

Appearing in Proceedings of the 13^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9. Copyright 2010 by the authors.



(a) $\nu = \frac{1}{2}$ (b) $\nu = 5$ (c) $\nu = 1000$

Figure 1: Three operation modes for a NLGBN unit. The black solid line shows the zero mean distribution (i.e., $y = 0$), the red dashed line shows a pre-sigmoid mean of $+1$ and the blue dash-dot line shows a pre-sigmoid mean of -1. (a) Binary behavior resulting from small precision. (b) Roughly Gaussian behavior resulting from medium precision. (c) Deterministic behavior from large precision.

variables) only in layer $m\!=\!0$. We require that units in layer m have parents only in layer $m\!+\!1$. Within layer m, we denote the number of units as $K^{(m)}$ and index the units with k so that the k^{th} unit in layer m is denoted $u_k^{(m)}$. We use the notation $\boldsymbol{u}^{(m)}$ to refer to the vector of all $K^{(m)}$ units for layer m together. A binary $K^{(m-1)}\!\times\! K^{(m)}$ matrix $\boldsymbol{Z}^{(m)}$ specifies the edges from layer m to layer $m\!-\!1$, so that element $Z_{k,k'}^{(m)}\!=\!1$ iff there is an edge from unit $u_{k'}^{(m)}$ to unit $u_k^{(m-1)}$.

A unit's activation is determined by a weighted sum of its parents. The weights for layer m are denoted by a $K^{(m-1)}\!\times\! K^{(m)}$ real-valued matrix $\boldsymbol{W}^{(m)}$, so that the activations for the units in layer m can be written as $\boldsymbol{y}^{(m)}\!=\!(\boldsymbol{W}^{(m+1)}\!\odot\!\boldsymbol{Z}^{(m+1)})\boldsymbol{u}^{(m+1)}\!+\!\boldsymbol{\gamma}^{(m)}$, where $\boldsymbol{\gamma}^{(m)}$ is a $K^{(m)}$-dimensional vector of *bias weights* and the binary operator $\odot$ indicates the Hadamard product.

To achieve a wide range of possible behaviors, we use the *nonlinear Gaussian belief network* (NLGBN) (Frey, 1997; Frey and Hinton, 1999) framework. In the NLGBN, the distribution on $u_k^{(m)}$ arises from adding zero mean Gaussian noise with precision $\nu_k^{(m)}$ to the activation sum $y_k^{(m)}$. This noisy sum is transformed with a sigmoid function $\sigma(\cdot)$ to obtain the value of the unit. We modify the NLGBN slightly so that the sigmoid function is from the real line to $(-1,1)$, i.e., $\sigma:\mathbb{R}\!\rightarrow\!(-1,1)$, via $\sigma(x) = 2/(1+\exp\{x\})-1$. The distribution of $u_k^{(m)} \in (-1,1)$ given its parents is then

$$p(u_k^{(m)}|y_k^{(m)},\nu_k^{(m)}) = \frac{\exp\left\{-\frac{\nu_k^{(m)}}{2}\left[\sigma^{-1}(u_k^{(m)})-y_k^{(m)}\right]^2\right\}}{\sigma'(\sigma^{-1}(u_k^{(m)}))\sqrt{2\pi/\nu_k^{(m)}}}$$

where $\sigma'(x) = \frac{\mathrm{d}}{\mathrm{d}x}\sigma(x)$. As discussed by Frey (1997) and as shown in Fig 1, different choices of $\nu_k^{(m)}$ yield different belief unit behaviors, ranging from effectively discrete binary units to nonlinear continuous units.

3 Infinite Belief Networks

Conditioned on the number of layers M, the layer widths $K^{(m)}$, and the network structures $\boldsymbol{Z}^{(m)}$, inference in belief networks can be straightforwardly implemented using Markov chain Monte Carlo (Neal, 1992). Learning the depth, width, and structure, however, presents significant computational challenges. In this section, we present a novel nonparametric prior, the *cascading Indian buffet process*, for multi-layered belief networks that are both infinitely wide and infinitely deep. By using an infinite prior we avoid the need for the complex dimensionality-altering proposals that would otherwise be required during inference.

3.1 The Indian buffet process

Sec 2 used the binary matrix $\boldsymbol{Z}^{(m)}$ as a convenient way to represent the edges connecting layer m to layer $m\!-\!1$. We stated that $\boldsymbol{Z}^{(m)}$ was a finite $K^{(m-1)}\!\times\! K^{(m)}$ matrix. We can use the *Indian buffet process* (IBP) (Griffiths and Ghahramani, 2006) to allow this matrix to have an infinite number of columns. We assume the two-parameter IBP (Ghahramani et al., 2007), and use $\boldsymbol{Z}^{(m)}\!\sim\!\mathsf{IBP}(\alpha,\beta)$ to indicate that the matrix $\boldsymbol{Z}^{(m)}\in\{0,1\}^{K^{(m-1)}\times\infty}$ is drawn from an IBP with parameters $\alpha,\beta>0$. The eponymous metaphor for the IBP is a restaurant with an infinite number of dishes available. Each customer chooses a finite set of dishes to taste. The rows of the binary matrix correspond to customers, while the columns correspond to dishes. If the j^{th} customer tastes the k^{th} dish, then $Z_{j,k}\!=\!1$. Otherwise, $Z_{j,k}\!=\!0$. The first customer to enter the restaurant samples a number of dishes that is Poisson-distributed with parameter α. After that, when the j^{th} customer enters the restaurant, she selects dish k with probability $\eta_k\,/\,(j\!+\!\beta\!-\!1)$, where η_k is the number of previous customers that have tried the k^{th} dish. She then chooses some number of additional dishes to taste that is Poisson-distributed with parameter $\alpha\beta\,/\,(j\!+\!\beta\!-\!1)$. Even though each customer chooses dishes based on their popularity among the previous customers, the rows and columns of the resulting matrix $\boldsymbol{Z}^{(m)}$ are infinitely exchangeable.

As with the model of Wood et al. (2006), if the model in Sec 2 had only a single hidden layer, i.e., $M\!=\!1$, then the IBP could be used to make that layer infinitely wide. Without intra-layer connections, however, the hidden units are independent *a priori*. This "shallowness" is a strong assumption that weakens the model in practice. The explosion of recent literature on *deep belief networks* (see, e.g., Hinton et al. (2006); Hinton and Salakhutdinov (2006)) speaks to the empirical success of networks with more hidden structure.

3.2 The cascading Indian buffet process

To build a prior on belief networks that are unbounded in both width and depth, we use an IBP-like construction that results in an infinite sequence of binary matrices $\boldsymbol{Z}^{(0)}, \boldsymbol{Z}^{(1)}, \boldsymbol{Z}^{(2)}, \cdots$. The matrices in this sequence must inherit the useful sparsity properties of the IBP, with the constraint that the columns from $\boldsymbol{Z}^{(m-1)}$ correspond to the rows in $\boldsymbol{Z}^{(m)}$. We interpret each matrix $\boldsymbol{Z}^{(m)}$ as specifying the directed edge structure from the units in layer m to those in layer $m-1$, where both layers have a potentially-unbounded width.

We propose the cascading Indian buffet process (CIBP), which provides a prior with these properties. The CIBP extends the IBP as follows: each dish in the restaurant is also a customer in another Indian buffet process—i.e., the columns in one binary matrix correspond to the rows in another. The CIBP is infinitely exchangeable in the rows of matrix $\boldsymbol{Z}^{(0)}$. Each of the matrices in the recursion is exchangeable in its rows and columns—propagating a permutation through the matrices does not change the probability of the data.

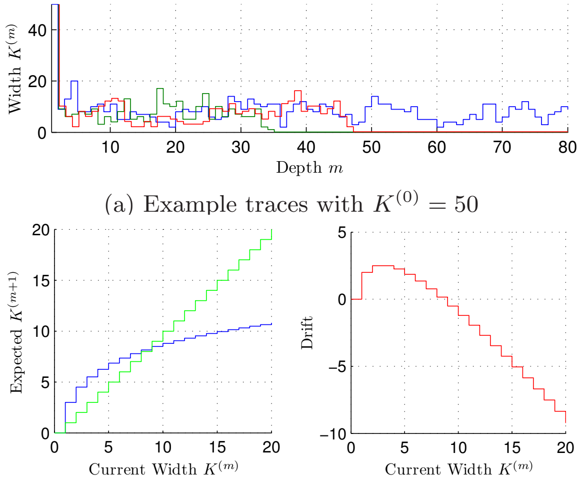
Surprisingly, if there are $K^{(0)}$ customers in the first restaurant, then for finite $K^{(0)}$, α, and β, the CIBP recursion terminates with probability one. In other words, at some point the customers do not taste any dishes, and deeper restaurants have neither dishes nor customers. Here we sketch the intuition behind this result. (See the supplementary materials for a proof.)

The CIBP constructs matrices in a sequence, starting with $m{=}0$. The number of nonzero columns in matrix $\boldsymbol{Z}^{(m+1)}$, $K^{(m+1)}$, is determined by $K^{(m)}$, the number of active nonzero columns in $\boldsymbol{Z}^{(m)}$. We require that for some matrix $\boldsymbol{Z}^{(m)}$, all columns are zero. We can therefore disregard the fact that the CIBP is a matrix-valued stochastic process and instead consider the Markov chain on the number of nonzero columns. Fig 2a shows three traces of such a Markov chain on $K^{(m)}$. If we define $\lambda(K;\alpha,\beta) = \alpha \sum_{k'=1}^{K} \frac{\beta}{k'+\beta-1}$, then the Markov chain has the transition distribution

$$p(K^{(m+1)} = k \,|\, K^{(m)}, \alpha, \beta) = \frac{1}{k!} \exp\left\{-\lambda(K^{(m)};\alpha,\beta)\right\} \lambda(K^{(m)};\alpha,\beta)^{k}, \quad (1)$$

which is a Poisson distribution with mean $\lambda(K^{(m)};\alpha,\beta)$. To show that the chain reaches the absorbing state $K^{(m)}{=}0$ with probability one, we must show that $K^{(m)}$ does not blow up to infinity.

In such a Markov chain, this requirement is equivalent to the chain having an equilibrium distribution when conditioned on nonabsorption (has a *quasi-stationary distribution*) (Seneta and Vere-Jones, 1966). For countably-infinite state spaces, a Markov chain has a (quasi-) stationary distribution if it is positive-recurrent, i.e., there is a finite expected time between consecutive visits to any state. Positive recurrency can be shown via the *Foster–Lyapunov stability criterion* (FLSC) (Fayolle et al., 2008). Satisfying the FLSC for the Markov chain with transitions given by Eqn 1 demonstrates that eventually the CIBP will reach a restaurant in which the customers try no new dishes. We do this by showing that if $K^{(m)}$ is large enough, then the expected $K^{(m+1)}$ is smaller than $K^{(m)}$. We use a *Lyapunov function* $\mathcal{L}(k) : \mathbb{N}^{+} \rightarrow \mathbb{R} > 0$, $\mathcal{L}(0){=}0$, with which we define the *drift function* as follows:



(a) Example traces with $K^{(0)} = 50$

(b) Expected $K^{(m+1)}$

(c) Drift

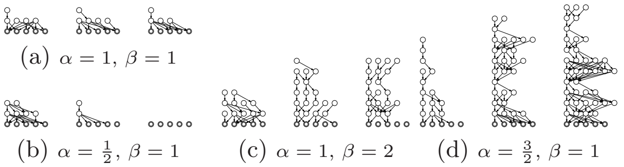
Figure 2: Properties of the Markov chain on layer width for the CIBP, with $\alpha = 3$, $\beta = 1$. Note that these values are illustrative and are not necessarily appropriate for a network structure. a) Example traces of a Markov chain on layer width, indexed by depth m. b) Expected $K^{(m+1)}$ as a function of $K^{(m)}$ is shown in blue. The Lyapunov function $\mathcal{L}(\cdot)$ is shown in green. c) The drift as a function of the current width $K^{(m)}$. This corresponds to the difference between the two lines in (a). Note that the drift becomes negative when the layer width is greater than eight.

$$\mathbb{E}_{k|K^{(m)}}[\mathcal{L}(k) - \mathcal{L}(K^{(m)})] = \sum_{k=1}^{\infty} p(K^{(m+1)} = k \,|\, K^{(m)})(\mathcal{L}(k) - \mathcal{L}(K^{(m)})).$$

The drift is the expected change in $\mathcal{L}(k)$ as a function of $K^{(m)}$. If there is a $K^{(m)}$ above which all drifts are negative, then the Markov chain satisfies the FLSC and is positive-recurrent. In the CIBP, this is satisfied for $\mathcal{L}(k) = k$. That the drift eventually will become negative can be seen by the fact that $\mathbb{E}_{k|K^{(m)}}[\mathcal{L}(k)] = \lambda(K^{(m)}\,;\,\alpha,\beta)$ is $O(\ln K^{(m)})$ and $\mathbb{E}_{k|K^{(m)}}[\mathcal{L}(K^{(m)})] = K^{(m)}$ is $O(K^{(m)})$. Figs 2b and 2c together provide a schematic illustration of this idea.

3.3 Unbounded priors on network structure

The CIBP can be used as a prior on the sequence $\boldsymbol{Z}^{(0)}, \boldsymbol{Z}^{(1)}, \boldsymbol{Z}^{(2)}, \cdots$ from Sec 2, to allow an infinite sequence of infinitely-wide hidden layers. As before, there are $K^{(0)}$ visible units. The edges between the



(a) $\alpha = 1$, $\beta = 1$

(b) $\alpha = \frac{1}{2}$, $\beta = 1$ (c) $\alpha = 1$, $\beta = 2$ (d) $\alpha = \frac{3}{2}$, $\beta = 1$

Figure 3: Samples from the CIBP prior (for four sets of α, β values) over network structures with five visible units.

first hidden layer and the visible layer are drawn according to the restaurant metaphor. This yields a finite number of units in the first hidden layer, denoted $K^{(1)}$ as before. These units become the visible units in another IBP-based network. While this recurses infinitely deep, only a finite number of units are ancestors of the visible units. If a unit has no ancestors, its activation is determined only by its bias. Fig 3 shows several samples from the prior for different parameterizations. Only connected units are shown.

A significant aspect of using an IBP-based prior is that it introduces a "rich-get-richer" behavior in the structure. The probability of a hidden unit acquiring a new outgoing edge increases with the current number of outgoing edges. While this is not desirable for all models, we feel it is a reasonable property of a belief network. Sharing of hidden variables in the belief network is what induces structure on the outputs. It seems appropriate that a hidden variable which is already important will likely become more important.

The parameters α and β govern the expected width and sparsity of the network at each level. The expected in-degree of each unit (number of parents) is α and the expected out-degree (number of children) is $K/\sum_{k=1}^{K} \frac{\beta}{\beta+k-1}$, for K units used in the layer below. These equations arise directly from the properties of the IBP described by Ghahramani et al. (2007). For clarity, we have presented the CIBP results with α and β fixed at all depths; however, this may be overly restrictive. For example, in an image recognition problem we would not expect the sparsity of edges mapping low-level features to pixels to be the same as that for high-level features to low-level features. To address this, we allow α and β to vary with depth, writing $\alpha^{(m)}$ and $\beta^{(m)}$. The CIBP terminates with probability one as long as there exists some finite upper bound for $\alpha^{(m)}$ and $\beta^{(m)}$ for all m. To ensure this, we place top-hat priors on $\alpha^{(m)}$ and $\beta^{(m)}$, which is vague but bounded.

3.4 Priors on other parameters

For other parameters in the model, we use priors that tie parameters together according to layer. We assume that the weights in layer m are drawn independently from Gaussian distributions with mean $\mu_w^{(m)}$ and precision $\rho_w^{(m)}$. We assume a similar layer-wise prior for biases $\boldsymbol{\gamma}^{(m)}$ with parameters $\mu_\gamma^{(m)}$ and $\rho_\gamma^{(m)}$. We use layer-wise gamma priors on the $\nu_k^{(m)}$, with parameters $a^{(m)}$ and $b^{(m)}$. We tie these prior parameters together with global normal-gamma hyperpriors for the weight and bias parameters, and gamma hyperpriors for the unit-activation precision parameters $\nu^{(m)}$.

4 Inference

We have so far described a prior on belief network structures and parameters, along with likelihood functions for unit activation. For inference, however, we must find the posterior distribution over the structure and the parameters of the network, having seen a set of data given by N D-dimensional vectors. We will assume that these data have been scaled to be in the range $(-1, 1)$. Our inference procedure assumes that these data have been generated by a belief network of some unknown width, depth, and structure. The first layer always has a width equal to the dimensionality of the data, i.e., $K^{(0)} = D$ and treat the data as activations of the visible units. This corresponds to pre-specifying the values of the visible units to be the data in N identically-structured networks. We denote this set of N visible layer units as $\{\boldsymbol{u}_n^{(0)}\}_{n=1}^N$.

We then wish to find the posterior distribution over: 1) the depth of the network; 2) the widths of the hidden layers; 3) the edge structure between layers; 4) the weights associated with the edges; 5) the biases of the units; 6) the activations of the hidden units that led to the data; 7) the values of the various hyperparameters. The posterior distribution over these unknowns is not analytically tractable, so we use Markov chain Monte Carlo (MCMC) to draw samples from it. To use MCMC, we instantiate these unknowns to particular values and then define a transition operator on this state that leaves the posterior distribution invariant. Under easily-satisfied conditions, the distribution over the current state of the Markov chain will evolve so as to be closer and closer to the distribution of interest.

From a technical standpoint, the trick with the CIBP (and other similar nonparametric Bayesian models) is that it does not actually define a prior on the width and depth of the network. Rather, it constructs a network with an infinite number of layers that each have an infinite number of units in such a way that only a finite number of units actually contribute to the observed data. In general, one would not expect that a distribution on infinite networks would yield tractable inference. However, in our construction, given the sequence $\boldsymbol{Z}^{(1)}, \boldsymbol{Z}^{(2)}, \cdots$, almost all of the infinite number of units are conditionally independent, unconnected to the data and therefore irrelevant to the posterior distribution. Due to this independence, the activations of these unconnected units trivially marginalize out of the model's joint distribution and we can restrict inference only to those units that are ancestors of the visible units. Of course, since this trivial marginalization only arises from the $\boldsymbol{Z}^{(m)}$ matrices, we must also have a distribution on infinite binary matrices that allows exact marginalization of all the uninstantiated edges. The row-wise and column-wise exchangeability properties of the IBP are what allows the use of infinite matrices. The bottom-up conditional structure of the CIBP allows an infinite number of these matrices.

To simplify notation, we will use $\boldsymbol{\Omega}$ for the aggregated state of the model variables. It is this aggregated state on which the Markov chain is defined. Given the hyperparameters, we write the joint distribution as

$$p(\boldsymbol{\Omega}) = \left(p(\boldsymbol{\gamma}^{(0)})\, p(\boldsymbol{\nu}^{(0)}) \prod_{k=1}^{K^{(0)}} \prod_{n=1}^{N} p(x_{k,n} \,|\, y_{k,n}^{(0)}, \nu_k^{(0)}) \right) \times \left(\prod_{m=1}^{\infty} p(\boldsymbol{W}^{(m)})\, p(\boldsymbol{\gamma}^{(m)})\, p(\boldsymbol{\nu}^{(m)}) \times \prod_{k=1}^{K^{(m)}} \prod_{n=1}^{N} p(u_{k,n}^{(m)} \,|\, y_{k,n}^{(m)}, \nu_k^{(m)}) \right). \quad (2)$$

Although this distribution involves several infinite sets, the aforementioned marginalization makes it possible to sample the relevant parts via MCMC. We use a sequence of transition operators that update subsets of the state, conditioned on the remainder, in such a way as to leave Eqn 2 invariant. We specifically note that conditioned on the binary matrices $\{\boldsymbol{Z}^{(m)}\}_{m=1}^{\infty}$, which define the structure of the network, inference becomes exactly as it would be in a finite belief network.

4.1 Updating hidden unit activations

Since we cannot easily integrate out the activations of the hidden units, we have included them as part of the MCMC state. Conditioned on the network structure, it is only necessary to sample the activations of the units that are ancestors of the visible units. Frey (1997) used slice sampling for the hidden unit states but we have had greater success with a specialized independence-chain variant of multiple-try Metropolis–Hastings (Liu et al., 2000). Our method proposes several (≈ 5) possible new unit activations from the prior imposed by its parents and selects among them (or rejects them all) according to the likelihood imposed by its children. As this operation can be executed in a vectorized manner with modern math libraries we have seen significantly better mixing performance by wall-clock time than with slice sampling.

4.2 Updating weights and biases

Given that a directed edge exists in the structure, we sample the posterior distribution over its weight. Conditioned on the rest of the model, the NLGBN provides a convenient Gaussian form for the distribution over weights so that we can Gibbs sample them from a conditional posterior Gaussian with parameters

$$\mu_{m,k,k'}^{\mathsf{w-post}} = \frac{\rho_w^{(m)} \mu_w^{(m)} + \nu_k^{(m-1)} \sum_n u_{n,k'}^{(m)} (\sigma^{-1}(u_k^{(m-1)}) - \xi_{n,k,k'}^{(m)})}{\rho_w^{(m)} + \nu_k^{(m-1)} \sum_n (u_{n,k'}^{(m)})^2}$$

$$\rho_{m,k,k'}^{\mathsf{w-post}} = \rho_w^{(m)} + \nu_k^{(m-1)} \sum_n (u_{n,k'}^{(m)})^2,$$

where

$$\xi_{n,k,k'}^{(m)} = \gamma_k^{(m-1)} + \sum_{k'' \neq k'} Z_{k,k''}^{(m)} W_{k,k''}^{(m)} u_{n,k''}^{(m)}. \quad (3)$$

The bias $\gamma_k^{(m)}$ can also be similarly sampled from a Gaussian conditional posterior distribution.

4.3 Updating activation variances

We use the NLGBN model to gain the ability to vary the mode of unit behaviors between discrete and continuous representations. This corresponds to sampling from the posterior distributions over the $\nu_k^{(m)}$. With a conjugate prior, the new value can be sampled from a gamma distribution with the following parameters:

$$a_{m,k}^{\nu\mathsf{-post}} = a_\nu^{(m)} + N/2 \quad (4)$$

$$b_{m,k}^{\nu\mathsf{-post}} = b_\nu^{(m)} + \frac{1}{2} \sum_{n=1}^{N} (\sigma^{-1}(u_{n,k}^{(m)}) - y_k^{(m)})^2. \quad (5)$$

4.4 Updating structure

To sample from the structure of the network—the sequence $\boldsymbol{Z}^{(1)}, \boldsymbol{Z}^{(2)}, \cdots$ in our construction—we must define an MCMC operator that adds and removes edges while leaving the posterior in Eqn 2 invariant. The procedure we use is similar to that proposed by Fox et al. (2009). When adding a layer, we must sample additional layer-wise model components. When introducing an edge, we must also sample its weight from the posterior distribution. If a new edge introduces a previously-unseen hidden unit, we must draw a bias for it and also draw its deeper-cascading connections from the posterior. Finally, we draw a top-down sample of the N new hidden unit activations from any unit we introduce. Effectively, we make a joint proposal for the edge and all relevant state that we previously did not have to store. We generate these proposals from the prior but define a Metropolis–Hastings rule that accepts them according to the posterior distribution.

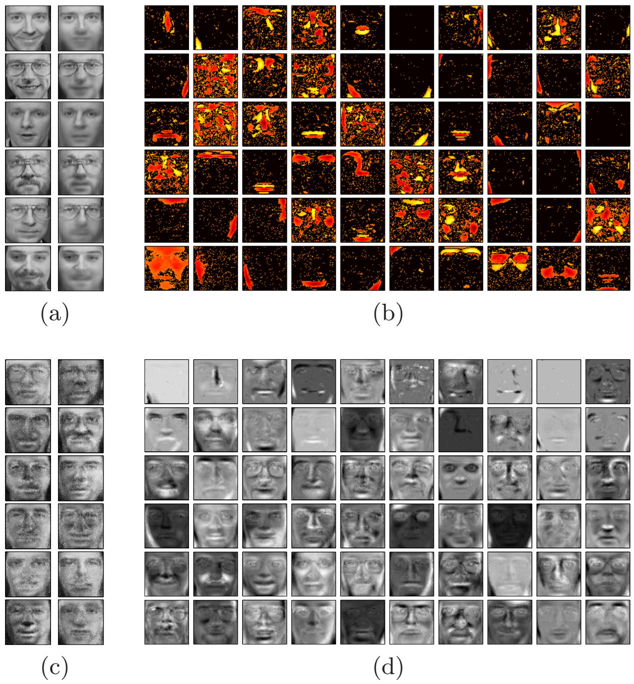


Figure 4: Olivetti faces a) Test images on the left, with reconstructed bottom halves on the right. b) Sixty features learned in the bottom layer, where black shows absence of an edge. Note the learning of sparse features corresponding to specific facial structures such as mouth shapes, noses and eyebrows. c) Raw predictive fantasies. d) Feature activations from individual units in the second hidden layer.

We iterate over each layer that connects to the visible units. Within each layer $m \geq 0$, we iterate over the connected units. Sampling the edges incident to the k^{th} unit in layer m has two phases. First, we iterate over each connected unit in layer $m+1$, indexed by k'. We calculate $\eta^{(m)}_{-k,k'}$, the number of nonzero entries in the k'^{th} column of $\boldsymbol{Z}^{(m+1)}$, excluding any entry in the k^{th} row. If $\eta^{(m)}_{-k,k'}$ is zero, we call the unit k' a *singleton* parent, to be dealt with in the second phase. If $\eta^{(m)}_{-k,k'}$ is nonzero, we introduce (or keep) the edge from unit $u^{(m+1)}_{k'}$ to $u^{(m)}_{k}$ with Bernoulli probability

$$p(Z^{(m+1)}_{k,k'}\!=\!1\,|\,\boldsymbol{\Omega}\backslash Z^{(m+1)}_{k,k'})=\frac{1}{\mathcal{Z}}\left(\frac{\eta^{(m)}_{-k,k'}}{K^{(m)}\!+\!\beta^{(m)}\!-\!1}\right)\prod_{n=1}^{N} p(u^{(m)}_{n,k}\,|\,Z^{(m+1)}_{k,k'}=1,\boldsymbol{\Omega}\backslash Z^{(m)}_{k,k'})$$

$$p(Z^{(m+1)}_{k,k'}\!=\!0\,|\,\boldsymbol{\Omega}\backslash Z^{(m+1)}_{k,k'})=\frac{1}{\mathcal{Z}}\left(1-\frac{\eta^{(m)}_{-k,k'}}{K^{(m)}\!+\!\beta^{(m)}\!-\!1}\right)\prod_{n=1}^{N} p(u^{(m)}_{n,k}\,|\,Z^{(m+1)}_{k,k'}=0,\boldsymbol{\Omega}\backslash Z^{(m+1)}_{k,k'}),$$

where $\mathcal{Z}$ is the appropriate normalization constant.

In the second phase, we consider deleting connections to singleton parents of unit k, or adding new singleton parents. We use Metropolis–Hastings with a birth/death process. If there are currently $K_\circ$ singleton parents, then with probability 1/2 we propose adding a new one by drawing it recursively from deeper layers, as above. We accept the proposal to insert a connection to this new parent unit with M–H ratio

$$\frac{\beta^{(m)}\!+\!K^{(m)}\!-\!1}{\alpha^{(m)}\beta^{(m)}(K_\circ\!+\!1)^2}\prod_{n=1}^{N}\frac{p(u^{(m)}_{n,k}\,|\,Z^{(m+1)}_{k,j}\!=\!1,\boldsymbol{\Omega}\backslash Z^{(m+1)}_{k,j})}{p(u^{(m)}_{n,k}\,|\,Z^{(m+1)}_{k,j}\!=\!0,\boldsymbol{\Omega}\backslash Z^{(m+1)}_{k,j})}.$$

If we do not propose to insert a unit and $K_\circ \geq 0$, then with probability 1/2 we select uniformly from among the singleton parents of unit k and propose removing the connection to it. We accept the proposal to remove the j^{th} one with M–H acceptance ratio given by

$$\frac{\alpha^{(m)}\beta^{(m)}K_\circ^2}{\beta^{(m)}\!+\!K^{(m)}\!-\!1}\prod_{n=1}^{N}\frac{p(u^{(m)}_{n,k}\,|\,Z^{(m+1)}_{k,j}\!=\!0,\boldsymbol{\Omega}\backslash Z^{(m+1)}_{k,j})}{p(u^{(m)}_{n,k}\,|\,Z^{(m+1)}_{k,j}\!=\!1,\boldsymbol{\Omega}\backslash Z^{(m+1)}_{k,j})}.$$

After these phases, chains of units that are not ancestors of the visible units can be discarded. Notably, this birth/death operator samples from the IBP posterior with a nontruncated equilibrium distribution, even without conjugacy. Unlike the stick-breaking approach of Teh et al. (2007), it allows use of the two-parameter IBP, which is important to this model.

5 Reconstructing Images

We applied our approach to three image data sets—the Olivetti faces, the MNIST digits and the Frey faces—and analyzed the structures that arose in the model posteriors. To assess the model, we constructed a missing-data problem using held-out images from each set. We removed the bottom halves of the test images and used the model to reconstruct the missing data, conditioned on the top half. Prediction itself was done by integrating out the parameters and structure via MCMC and averaging over predictive samples.

Olivetti Faces The Olivetti faces data (Samaria and Harter, 1994) are 400 64×64 grayscale images of the faces of 40 distinct subjects, which we divided randomly into 350 training and 50 test data. Fig 4a shows six bottom-half test set reconstructions on the right, compared to the ground truth on the left. Fig 4b shows a subset of sixty weight patterns from a posterior sample of the structure, with black indicating that no edge is present from that hidden unit to the visible unit (pixel). The algorithm is clearly assigning hidden units to specific and interpretable features, such as mouth shapes, facial hair, and skin tone. Fig 4c shows ten pure fantasies from the model, easily generated in a directed acyclic belief network. Fig 4d shows the result of activating individual units in the second

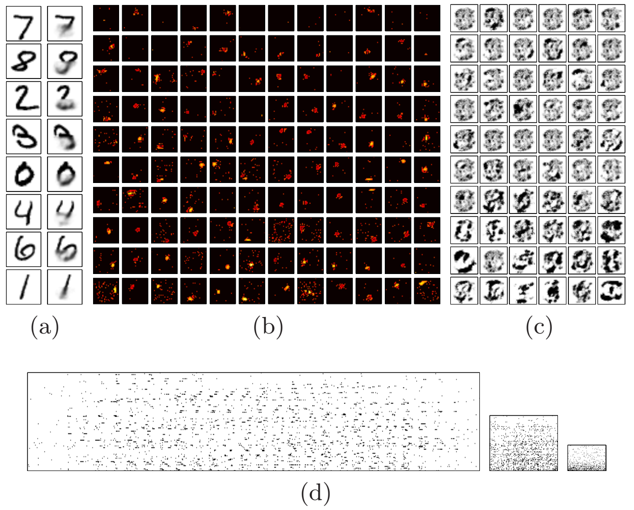


Figure 5: MNIST Digits a) Eight pairs of test image reconstructions, with the bottom half of each digit missing. The truth is the left image in each pair. b) 120 features learned in the bottom layer, where black indicates that no edge exists. c) Activations in pixel space resulting from activating individual units in the deepest layer. d) Samples from the posterior of $\boldsymbol{Z}^{(0)}$, $\boldsymbol{Z}^{(1)}$ and $\boldsymbol{Z}^{(2)}$ (transposed).

hidden layer, while keeping the rest unactivated, and propagating the activations down to the visible pixels. This provides an idea of the image space spanned by the principal components of these deeper units. A typical posterior network structure had three hidden layers, with approximately seventy units in each layer.

MNIST Digits We used a subset of the MNIST handwritten digit data (LeCun et al., 1998) for training: 50 28 × 28 examples of each of the ten digits, with ten more examples of each digit held out for testing. The inferred lower-level features are extremely sparse, as shown in Fig 5b, and the deeper units are simply activating sets of blobs at the pixel level. This is shown also by activating individual units at the deepest layer, as shown in Fig 5c. Test reconstructions are in Fig 5a. A typical network had three hidden layers, with roughly 120, 100, and 70 units in each one. Fig 5d shows typical binary matrices $\boldsymbol{Z}^{(0)}$, $\boldsymbol{Z}^{(1)}$, and $\boldsymbol{Z}^{(2)}$.

Frey Faces The Frey faces data[1] are 1965 20 × 28 grayscale video frames of a single face with different expressions. We randomly selected 1865 training images and 100 test images. While typical posterior samples of the network again typically used three hidden layers, the networks for these data tended to be much wider and more densely connected. In the bottom layer, as shown in Fig 6b, a typical hidden unit connects to many pixels. We attribute this to global correlation effects since all images come from a single person. Typical layer widths were around 260, 120, and 35 units.

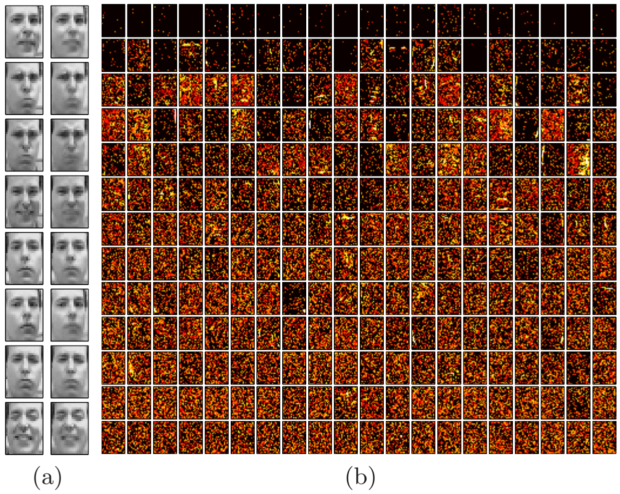


Figure 6: Frey faces a) Eight pairs of test reconstructions, with the bottom half of each face missing. The truth is the left image in each pair. b) 260 features learned in the bottom layer, where black indicates that no edge exists.

In the experiments, our MCMC sampler appeared to mix well and begins to find reasonable reconstructions after a few hours of CPU time. Note that the learned sparse connection patterns in $\boldsymbol{Z}^{(0)}$ varied from local (MNIST), through intermediate (Olivetti) to global (Frey), despite identical IBP hyperpriors. This suggests that flexible priors on structures are needed to adequately capture the statistics of different data sets.

6 Discussion

This paper unites two areas of research—nonparametric Bayesian methods and deep belief networks—to provide a novel nonparametric perspective on the general problem of learning the structure of directed deep belief networks with hidden units.

We addressed three issues surrounding deep belief networks. First, we inferred appropriate local representations with units varying from discrete binary to nonlinear continuous behavior. Second, we provided a way for a deep belief network to contain an arbitrary number of hidden units arranged in an arbitrary number of layers. Third, we presented a method for inferring the graph structure of a directed deep belief network. To achieve this, we introduced a *cascading* extension to the Indian buffet process and proved convergence properties that make it useful as a Bayesian prior distribution for a sequence of infinite binary matrices.

This work can be viewed as an infinite multilayer generalization of the density network (MacKay, 1995), and also as part of a more general literature of learning structure in probabilistic networks. With a few exceptions (e.g., Ramachandran and Mooney (1998); Friedman (1998); Elidan et al. (2000); Beal and Ghahramani (2006)), most previous work on learning the

[1] `http://www.cs.toronto.edu/~roweis/data.html`

structure of belief networks has focused on the case where all units are observed (Buntine, 1991; Heckerman et al., 1995; Friedman and Koller, 2003; Koivisto and Sood, 2004). Our framework not only allows for an unbounded number of hidden units, but couples the model for the number of units with a nonparametric model for the structure of the network. Rather than comparing structures by evaluating marginal likelihoods, we do inference in a single model with an unbounded number of units and layers, thereby learning effective model complexity. This approach is more appealing both computationally and philosophically.

There are a variety of future research directions arising from the model we have presented here. As we have presented it, we do not expect that our unsupervised, MCMC-based inference scheme will be competitive on supervised tasks with extensively-tuned discriminative models based on variants of maximum-likelihood learning. However, we believe that our model can inform choices for the network depth, layer size, and edge structure in such networks, and will inspire further research into flexible, nonparametric network models.

Acknowledgements

The authors thank Brendan Frey, David MacKay, Iain Murray and Radford Neal for valuable discussions. RPA is funded by the Canadian Institute for Advanced Research. This work was supported in part by the Center for Intelligent Information Retrieval, in part by CIA, NSA and NSF under NSF grant #IIS-0326249, and in part by subcontract #B582467 from Lawrence Livermore National Security, LLC, prime contractor to DOE/NNSA contract #DE-AC52-07NA27344. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

References

M. J. Beal and Z. Ghahramani. Variational Bayesian learning of directed graphical models with hidden variables. *Bayesian Analysis*, 1(4):793–832, 2006.

W. Buntine. Theory refinement on Bayesian networks. In *7th Conference on Uncertainty in Artificial Intelligence*, 1991.

A. C. Courville, D. Eck, and Y. Bengio. An infinite factor model hierarchy via a noisy-or mechanism. In *Advances in Neural Information Processing Systems 22*, 2009.

G. Elidan, N. Lotner, N. Friedman, and D. Koller. Discovering hidden variables. In *Advances in Neural Information Processing Systems 13*, 2000.

G. Fayolle, V. A. Malyshev, and M. V. Menshikov. *Topics in the Constructive Theory of Countable Markov Chains.* Cambridge University Press, Cambridge, UK, 2008.

E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky. Sharing features among dynamical systems with beta processes. In *Advances in Neural Information Processing Systems 22*, 2009.

B. J. Frey. Continuous sigmoidal belief networks trained using slice sampling. In *Advances in Neural Information Processing Systems 9*, 1997.

B. J. Frey and G. E. Hinton. Variational learning in nonlinear Gaussian belief networks. *Neural Computation*, 11 (1):193–213, 1999.

N. Friedman. The Bayesian structural EM algorithm. In *14th Conference on Uncertainty in Artificial Intelligence*, 1998.

N. Friedman and D. Koller. Being Bayesian about network structure. *Machine Learning*, 50(1-2):95–125, 2003.

Z. Ghahramani, T. L. Griffiths, and P. Sollich. Bayesian nonparametric latent feature models. In *Bayesian Statistics 8*, pages 201–226. 2007.

T. L. Griffiths and Z. Ghahramani. Infinite latent feature models and the Indian buffet process. In *Advances in Neural Information Processing Systems 18*, 2006.

D. Heckerman, D. Geiger, and D. M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3):197–243, 1995.

G. E. Hinton and R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313 (5786):504–507, 2006.

G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18 (7):1527–1554, July 2006.

M. Koivisto and K. Sood. Exact Bayesian structure discovery in Bayesian networks. *Journal of Machine Learning Research*, 5:549–573, 2004.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

J. S. Liu, F. Liang, and W. H. Wong. The use of multiple-try method and local optimization in Metropolis sampling. *Journal of the American Statistical Association*, 95(449):121–134, 2000.

D. J. C. MacKay. Bayesian neural networks and density networks. *Nuclear Instruments and Methods in Physics Research, Section A*, 354(1):73–80, 1995.

R. M. Neal. Connectionist learning in belief networks. *Artificial Intelligence*, 56:71–113, July 1992.

J. Pearl. *Probabalistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA, 1988.

S. Ramachandran and R. J. Mooney. Theory refinement of Bayesian networks with hidden variables. In *15th International Conference on Machine Learning*, 1998.

F. S. Samaria and A. C. Harter. Parameterisation of a stochastic model for human face identification. In *2nd IEEE Workshop on Applications of Comp. Vision*, 1994.

E. Seneta and D. Vere-Jones. On quasi-stationary distributions in discrete-time Markov chains with a denumerable infinity of states. *Journal of Applied Probability*, 3 (2):403–434, December 1966.

Y.-W. Teh, D. Görür, and Z. Ghahramani. Stick-breaking construction for the Indian buffet process. In *11th International Conference on Artificial Intelligence and Statistics*, 2007.

F. Wood, T. L. Griffiths, and Z. Ghahramani. A non-parametric Bayesian method for inferring hidden causes. In *22nd Conference on Uncertainty in Artificial Intelligence*, 2006.