Learning the structure of
deep sparse directed graphical models

Ryan P. Adams[1], Hanna M. Wallach[2], and Zoubin Ghahramani[3]

[1]University of Toronto
[2]University of Massachusetts at Amherst
[3]University of Cambridge

Machine Learning Summer School
August 2009

Motivation

- Present some recent research on graphical model structure learning...

- ...related to deep belief networks...

- ...which uses Markov chain Monte Carlo inference...

- ...in a non-parametric Bayesian model.

Deep networks

There is a great deal of interest on "deep belief networks".

Deep belief nets are probabilistic generative models that are composed of multiple layers of stochastic, latent variables. The latent variables typically have binary values and are often called hidden units or feature detectors. The top two layers have undirected, symmetric connections between them and form an associative memory. The lower layers receive top-down, directed connections from the layer above. The states of the units in the lowest layer represent a data vector.

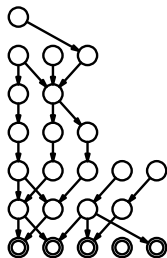Geoffrey E. Hinton (2009) Scholarpedia.

Deep networks

Questions:

- ► How many layers should there be?
- ► How wide should each layer be?
- ► What sorts of units?

Goal: To learn the structure of a deep network.

Approach: A *nonparametric Bayesian* method that learns the structure of a layered *directed* deep belief network.

Layered directed deep networks

$$p(\mathbf{x}) = \prod_{i=1}^{K} p(x_i | \mathbf{x}_{\pi_i})$$

Where $\mathbf{x} = (x_1, \ldots, x_K)$ and π_i are the parents of node i.

aka Bayesian networks, probabilistic directed graphical models.

Assume a layered graph structure.

How many layers? How wide should each layer be?

Priors over graph structures

Let $z_{ij}^{(m)} = 1$ mean that $j \in \pi_i$, that is, node j is a parent of node i in layer m.

If we specify a sequence of matrices $Z^{(0)}$, $Z^{(1)}$, $Z^{(2)}$, ... we have defined the layered graph structure.

Previous work

A Non-Parametric Bayesian Method for Inferring Hidden Causes
(Wood, Griffiths, Ghahramani, 2006)

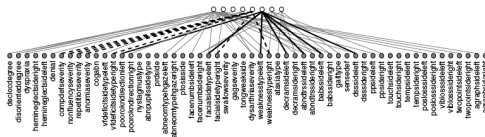Inferring stroke localization from patient symptoms:

Figure 6: Causal structure with highest posterior probability. Two grouping of signs are highlighted. In solid black, we find a grouping of poor optokinetic nystagmus, lack of facial control, weakness, decreased rapid alternating movements, abnormal deep tendon reflexes, Babinski sign, and double simultaneous stimulation neglect, all on the left side, consistent with a right frontal/parietal infarct. In dashed black, we find a grouping of comprehension deficit, non-fluency, repetition, anomia, visual field deficit, facial weakness, and general weakness, with the latter three on the right side, generally consistent with a left temporal infarct.

Y - binary latent factors (diseases, k)
Z - graph structure ($\sim$ Indian Buffet Process (IBP))
X - observed binary features (56 symptoms, s)
Noisy-or: $P(x_{sp} = 1|Z, Y, \lambda, \epsilon) = 1 - (1 - \lambda)^{\sum_k z_{sk} y_{kp}}(1 - \epsilon)$

The IBP defines a distribution on sparse binary matrices with a countably infinite number of columns.

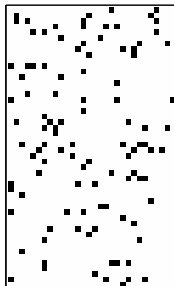Graphical models with an unbounded number of hidden units

Preview of the Indian buffet process
From finite to infinite matrices

$z_{ik} = 1$ means object i has feature k:

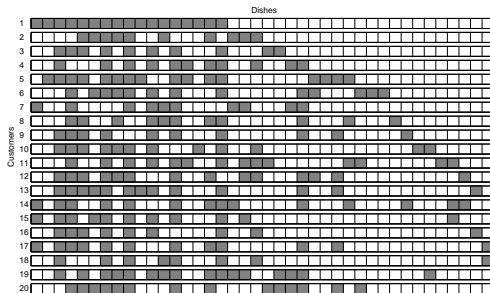$$z_{ik} \sim \text{Bernoulli}(\theta_k)$$

$$\theta_k \sim \text{Beta}(\alpha/K, 1)$$

▶ Note that $P(z_{ik} = 1|\alpha) = E(\theta_k) = \frac{\alpha/K}{\alpha/K+1}$, so as K grows larger the matrix gets sparser.

▶ So if Z is $N \times K$, the expected number of nonzero entries is $N\alpha/(1 + \alpha/K) < N\alpha$.

▶ Even in the $K \to \infty$ limit, the matrix is expected to have a finite number of non-zero entries.

▶ Two parameter extension $\theta_k \sim \text{Beta}(\alpha\beta/K, \beta)$

Indian buffet process

- ▶ First customer starts at the left of the buffet, and takes a serving from each dish, stopping after a Poisson(α) number of dishes as his plate becomes overburdened.
- ▶ The nth customer moves along the buffet, sampling dishes in proportion to their popularity, serving himself dish k with probability m_k/n, and trying Poisson(α/n) new dishes.
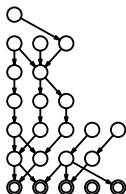- ▶ The customer-dish matrix is the feature matrix, Z.

Cascading Indian buffet process
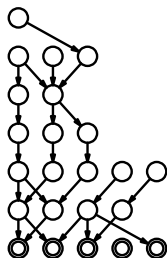
Start with $K^{(0)}$ rows (visible units)

- $Z^{(0)} \sim IBP(\alpha, \beta)$ with $K^{(0)}$ rows and $K^{(1)}$ non-zero columns
- $Z^{(1)} \sim IBP(\alpha, \beta)$ with $K^{(1)}$ rows and $K^{(2)}$ non-zero columns
- $Z^{(2)} \sim IBP(\alpha, \beta)$ with $K^{(2)}$ rows and $K^{(3)}$ non-zero columns
- ...

This defines a sequences of infinite sparse binary matrices.

Properties of the Cascading IBP

$$Z^{(m)} \sim IBP(\alpha, \beta) \quad \text{for } m = 0, 1, 2, \dots$$

- ▶ The expected in-degree of each unit (number of parents) is α.
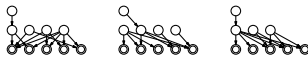- ▶ The expected out-degree of each unit in m (number of children) is

$$c(\beta, m) = 1 + \frac{K^{(m-1)} - 1}{1 + \beta}$$

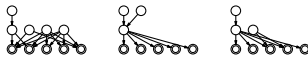 Note that $\lim_{\beta \to 0} c(\beta, m) = K^{(m-1)}$ and $\lim_{\beta \to \infty} c(\beta, m) = 1$.

- ▶ Hidden units are exchangeable at each layer.
- ▶ **Theorem**: For $K^{(m)} \in \mathbb{N}$, $0 < \alpha < \infty$, $0 < \beta < \infty$, the sequence of $K^{(m)}$ defined by the CIBP reaches the absorption state 0, with probability one, i.e. $\lim_{m \to \infty} p(K^{(m)} = 0) = 1$.

Samples from the prior over structures
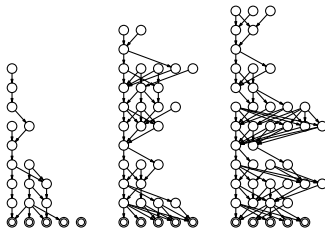
Samples from the CIBP prior starting from five visible units.

What kinds of units?

We want a model that is flexible enough to learn what types of unit it needs, ranging from binary to linear-Gaussian.

This idea was explored in Nonlinear Gaussian belief networks (NLGBNs) by (Frey and Hinton, 1999).

Let $\mathbf{u}^{(m)}$ be the activity of units in layer m.

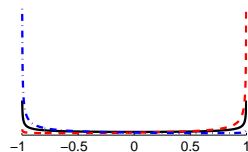$$\mathbf{y}^{(m)} = (W^{(m+1)} \odot Z^{(m+1)})\mathbf{u}^{(m+1)} + \boldsymbol{\gamma}^{(m)}$$

where W is a weight matrix, $\boldsymbol{\gamma}$ is a bias vector and $\odot$ is Hadamard (elementwise) product.
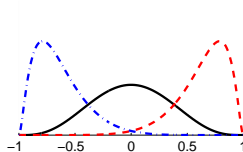
$$u_k^{(m)} = \sigma(y_k^{(m)} + \epsilon_k^{(m)})$$

σ is a sigmoid function and noise $\epsilon_k^{(m)} \sim \mathcal{N}(0, \frac{1}{\nu_k^{(m)}})$ has precision $\nu_k^{(m)}$.
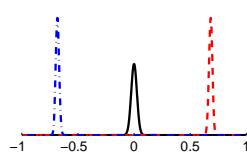
NLGBN units

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

Three modes of operation for the NLGBN unit. The black solid line shows the zero mean distribution, 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 from small precision.

(b) Roughly Gaussian behavior from medium precision.

(c) Deterministic behavior from large precision.
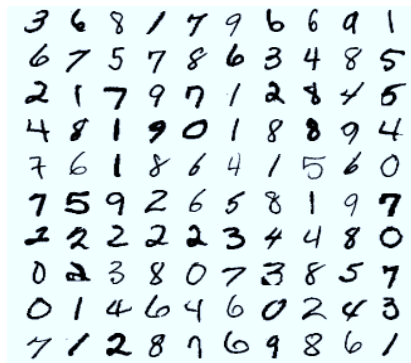
Inference
using Markov chain Monte Carlo

$$W \sim \mathcal{N} \quad \boldsymbol{\gamma} \sim \mathcal{N} \quad \alpha \sim \mathcal{G} \quad \beta \sim \mathcal{G} \quad \nu \sim \mathcal{G}$$

We design an MCMC scheme to sample from the posterior:

$$p(\{Z^{(m)}, W^{(m)}\}_{m=1}^{\infty}, \{\boldsymbol{\gamma}^{(m)}, \boldsymbol{\nu}^{(m)}\}_{m=0}^{\infty}, \{\{\mathbf{u}_n^{(m)}\}_{m=1}^{\infty}\}_{n=1}^{N} | \{\mathbf{x}_n\}_{n=1}^{N})$$

- ▶ $\mathbf{u}$ - slice sample
- ▶ W and γ - Gibbs
- ▶ ν - Gibbs
- ▶ Z - Gibbs (cf Algorithm 8 of CRPs)

Experiments on MNIST data

Small subset
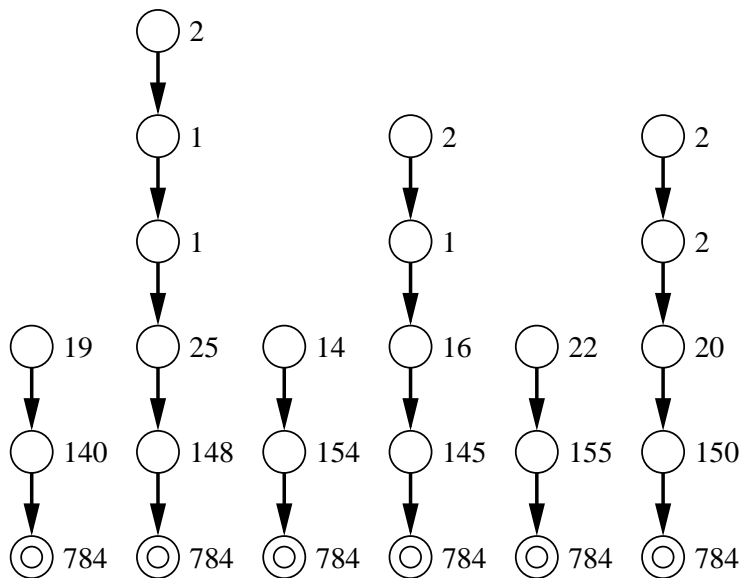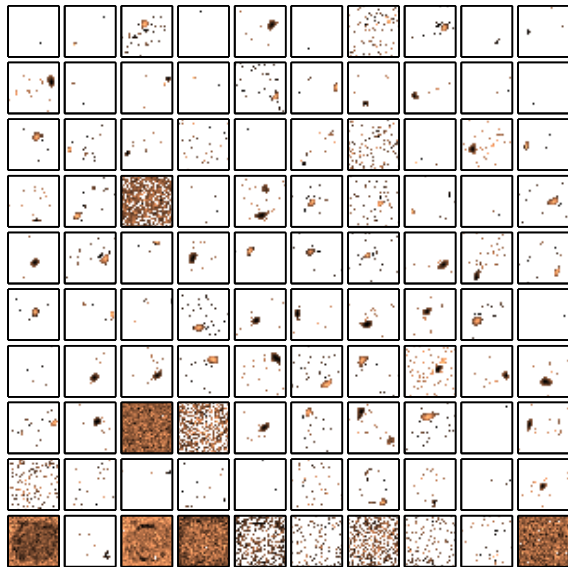28×28 pixels
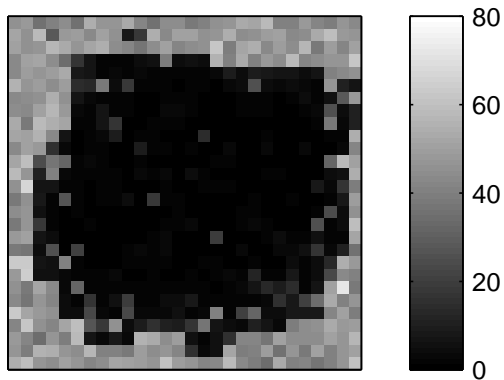100 images (10 from each class)

Samples from Posterior over Structures

First-Layer Features

Visible Unit Precisions

Summary

This work provides an initial attempt at addressing three issues with layered belief networks.

- ▶ It provides a way to learn belief networks that contain an arbitrary number of hidden units with nontrivial joint distributions due to a deep structure.
- ▶ It allows the units to have different operating regimes and infer appropriate local representations ranging from discrete binary to nonlinear continuous behavior.
- ▶ It provides a way to infer the appropriate directed graph structure of a layered network.

Initial work... many open questions!

Ryan P. Adams and Hanna M. Wallach

Overall Summary

- Graphical models provide a powerful and intuitive framework for modelling and inference.
- Directed, undirected and factor graphs.
- Inference by message passing.
- Parameter and structure learning.
- A recent bit of research on structure learning.

Thanks!

Questions