Bayesian Graphical Models

Steffen Lauritzen, University of Oxford

Graphical Models and Inference, Lecture 16, Michaelmas Term 2009

December 4, 2009
Parameter $\theta$, data $X = x$, likelihood

$$L(\theta \mid x) \propto p(x \mid \theta).$$

Express knowledge about $\theta$ through prior distribution $\pi$ on $\theta$. Inference about $\theta$ from $x$ is then represented through posterior distribution $\pi^*(\theta) = p(\theta \mid x)$. Then, from Bayes’ formula

$$\pi^*(\theta) = p(x \mid \theta) \pi(\theta) / p(x) \propto L(\theta \mid x) \pi(\theta)$$

so the likelihood function is equal to the density of the posterior w.r.t. the prior modulo a constant.
Represent statistical models as *Bayesian networks with parameters included as nodes*, i.e. for expressions as

\[ p(x_v \mid x_{pa(v)}, \theta_v) \]

*include \( \theta_v \) as additional parent of \( v \).* In addition, represent data explicitly in network using *plates.*

Then *Bayesian inference about \( \theta \) can in principle be calculated by probability propagation* as in general Bayesian networks.

This is *true for \( \theta_v \) discrete.* For \( \theta \) continuous, we must develop other computational techniques.
Chest clinic example with parameters and plate indicating repeated cases.
As for a naive Bayes expert system, just let $D = \theta$ and $X_i = F_i$ represent data.

Then $\pi^*(\theta) = P(\theta \mid X_1 = x_1, \ldots, X_m = X_m)$ is found by standard updating, using probability propagation if $\theta$ is discrete.
Bernoulli experiments

Data $X_1 = x_1, \ldots, X_n = x_n$ independent and Bernoulli distributed with parameter $\theta$, i.e.

$$P(X_i = 1 | \theta) = 1 - P(X_i = 0) = \theta.$$

Represent as a Bayesian network with $\theta$ as only parent to all nodes $x_i, i = 1, \ldots, n$. Use a beta prior:

$$\pi(\theta | a, b) \propto \theta^{a-1}(1 - \theta)^{b-1}.$$  

If we let $x = \sum x_i$, we get the posterior:

$$\pi^*(\theta) \propto \theta^x(1 - \theta)^{n-x}\theta^{a-1}(1 - \theta)^{b-1}$$

$$= \theta^{x+a-1}(1 - \theta)^{n-x+b-1}$$

So the posterior is also beta with parameters $(a + x, b + n - x)$.  

Steffen Lauritzen, University of Oxford  
Bayesian Graphical Models
Linear regression

```
model {
    for (i in 1 : N) {
        Y[i] ~ dnorm(mu[i], tau)
        mu[i] <- alpha + beta * (x[i] - xbar)
    }
    tau ~ dgamma(0.001, 0.001)
    sigma <- 1 / sqrt(tau)
    alpha ~ dnorm(0.0, 1.0E-6)
    beta ~ dnorm(0.0, 1.0E-6)
}
```
Gamma model for pumpdata

Failure of 10 power plant pumps.

Bayesian inference
Bayesian graphical models
Markov chain Monte Carlo methods
Simple examples
WinBUGS examples

for(i IN 1 : N)
beta alpha
x[i]
t[i]
theta[i]
lambda[i]

alpha
beta
theta[i]
t[i]
lambda[i]
x[i]

for(i IN 1 : N)
The number of failures $X_i$ is assumed to follow a Poisson distribution with parameter $\theta_i t_i$, $i = 1, \ldots, 10$ where $\theta_i$ is the failure rate for pump $i$ and $t_i$ is the length of operation time of the pump (in 1000s of hours). The data are shown below.

<table>
<thead>
<tr>
<th>Pump $t_i$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_i$</td>
<td>94.5</td>
<td>15.7</td>
<td>62.9</td>
<td>126</td>
<td>5.24</td>
<td>31.4</td>
<td>1.05</td>
<td>1.05</td>
<td>2.01</td>
<td>10.5</td>
</tr>
<tr>
<td>$x_i$</td>
<td>5</td>
<td>1</td>
<td>5</td>
<td>14</td>
<td>3</td>
<td>19</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>22</td>
</tr>
</tbody>
</table>

A gamma prior distribution is adopted for the failure rates:

$\theta_i \sim \Gamma(\alpha, \beta), i = 1, \ldots, 10$
With suitable priors the program becomes

```plaintext
model
{
  for (i in 1 : N) {
    theta[i] ~ dgamma(alpha, beta)
    lambda[i] <- theta[i] * t[i]
    x[i] ~ dpois(lambda[i])
  }
  alpha ~ dexp(1)
  beta ~ dgamma(0.1, 1.0)
}
```
Growth of 30 young rats.
Description of rat data

30 young rats have weights measured weekly for five weeks. The observations $Y_{ij}$ are the weights of rat $i$ measured at age $x_j$. The model is essentially a random effects linear growth curve:

$$Y_{ij} \sim \mathcal{N}(\alpha_i + \beta_i(x_j - \bar{x}), \tau_c^{-1})$$

and

$$\alpha_i \sim \mathcal{N}(\alpha_c, \tau_\alpha^{-1}), \quad \beta_i \sim \mathcal{N}(\beta_c, \tau_\beta^{-1})$$

where $\bar{x} = 22$, and $\tau$ represents the precision (inverse variance) of a normal distribution. Interest particularly focuses on the intercept at zero time (birth), denoted $\alpha_0 = \alpha_c - \beta_c \bar{x}$. 
When exact computation is infeasible, Markov chain Monte Carlo (MCMC) methods are used. An MCMC method for the target distribution $\pi^*$ on $X = X_V$ constructs a Markov chain $X^0, X^1, \ldots, X^k, \ldots$ with $\pi^*$ as equilibrium distribution. For the method to be useful, $\pi^*$ must be the unique equilibrium, and the Markov chain must be ergodic so that for all relevant $A$

$$
\pi^*(A) = \lim_{n \to \infty} \pi_n^*(A) = \lim_{n \to \infty} \frac{1}{n} \sum_{i=m+1}^{m+n} \chi_A(X^i)
$$

where $\chi_A$ is the indicator function of the set $A$. 

A simple MCMC method is made as follows.

1. Enumerate \( V = \{1, 2, \ldots, |V|\} \)
2. choose starting value \( x^0 = x_1^0, \ldots, x_{|V|}^0 \).
3. Update now \( x^0 \) to \( x^1 \) by replacing \( x_i^0 \) with \( x_i^1 \) for \( i = 1, \ldots, |V| \), where \( x_i^1 \) is chosen from ‘the full conditionals’

\[
\pi^*(X_i | x_1^1, \ldots, x_{i-1}^1, x_{i+1}^0, \ldots x_{|V|}^0).
\]

4. Continue similarly to update \( x^k \) to \( x^{k+1} \) and so on.
Properties of Gibbs sampler

With positive joint target density \( \pi^*(x) > 0 \), the Gibbs sampler is ergodic with \( \pi^* \) as the unique equilibrium. In this case the distribution of \( X^n \) converges to \( \pi^* \) for \( n \) tending to infinity.

Note that if the target is the conditional distribution

\[
\pi^*(x_A) = f(x_A \mid X_{V \setminus A} = x^*_{V \setminus A}),
\]

only sites in \( A \) should be updated:

The full conditionals of the conditional distribution are unchanged for unobserved sites.
For a directed graphical model, the density of full conditional distributions are:

\[
f(x_i \mid x_{V \setminus i}) \propto \prod_{v \in V} f(x_v \mid x_{pa(v)}) \\
\propto f(x_i \mid x_{pa(i)}) \prod_{v \in ch(i)} f(x_v \mid x_{pa(v)}) \\
= f(x_i \mid x_{bl(i)}),
\]

where \(bl(i)\) is the \textit{Markov blanket} of node \(i\):

\[
bl(i) = pa(i) \cup ch(i) \cup \{ \bigcup_{v \in ch(i)} pa(v) \setminus \{i\} \}.
\]

Note that \textit{the Markov blanket is just the neighbours of \(i\) in the moral graph}: \(bl(i) = ne^m(i)\).
There are many ways of sampling from a density $f$ which is known up to normalization, i.e. $f(x) \propto h(x)$. One uses an envelope $g(x) \geq Mh(x)$, where $g(x)$ is a known density and then proceeding as follows:

1. Choose $X = x$ from distribution with density $g$.
2. Choose $U = u$ uniform on the unit interval.
3. If $u > Mh(x)/g(x)$, then reject $x$ and repeat step 1, else return $x$.

The value returned will have density $f$. 