Bayesian Graphical Models

Steffen Lauritzen, University of Oxford

Graphical Models and Inference, Lectures 15 and 16, Michaelmas Term 2011

December 1, 2011

Parameter θ, data $X = x$, likelihood

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

Express knowledge about θ through *prior distribution* π on θ. Inference about θ 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.

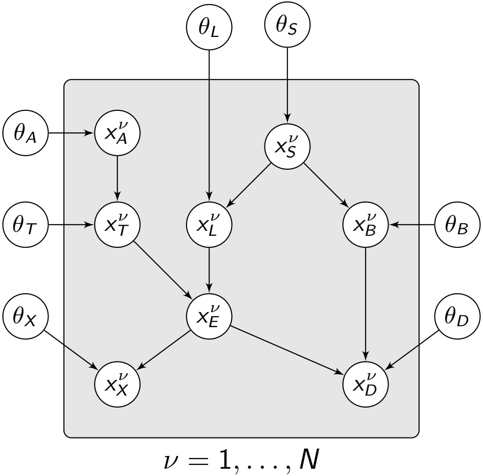
Represent statistical models as *Bayesian networks with parameters included as nodes,* i.e. for expressions as

$$p(x_v \mid x_{\mathrm{pa}(v)}, \theta_v)$$

include θ_v as additional parent of v. In addition, represent data explicitly in network using *plates.*

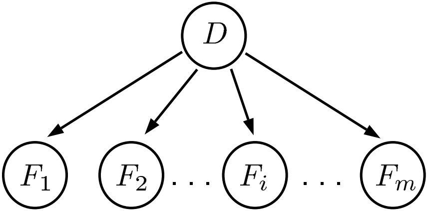
Then *Bayesian inference about θ can* in principle *be calculated by probability propagation* as in general Bayesian networks.

This is *true for θ_v discrete*. For θ continuous, we must develop other computational techniques.



Chest clinic with parameters and plate indicating repeated cases.

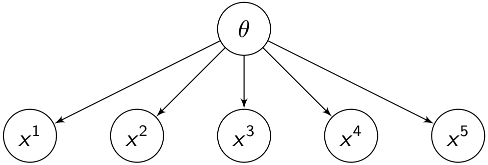
Standard repeated samples

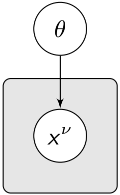


As for a naive Bayes expert system, just let $D = \theta$ and $X_i = F_i$ represent data.

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

Simple sampling represented with a plate





$\nu = 1, \ldots, 5$

Bernoulli experiments

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

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

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

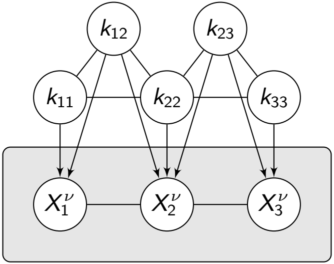
$$\pi(\theta \mid a, b) \propto \theta^{a-1}(1-\theta)^{b-1}.$$

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

$$\begin{aligned} \pi^*(\theta) &\propto \theta^x(1-\theta)^{n-x}\theta^{a-1}(1-\theta)^{b-1} \\ &= \theta^{x+a-1}(1-\theta)^{n-x+b-1} \end{aligned}$$

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

Bayesian variant of simple Gaussian graphical model



$\nu = 1, \ldots, N$

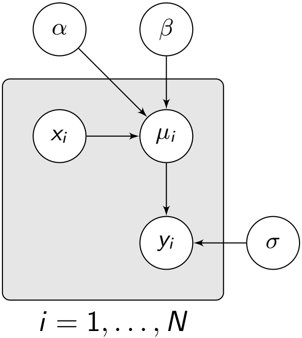
Parameters and repeated observations must be explicitly represented in the Bayesian model for $X_1 \perp\!\!\!\perp X_2 \mid X_3, K$. Here K follows a so-called hyper Markov prior, with further independence relations among the elements of K.

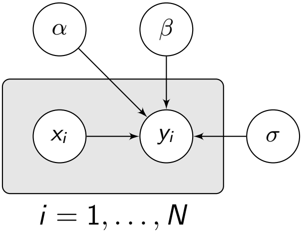
Linear regression

For the linear regression model

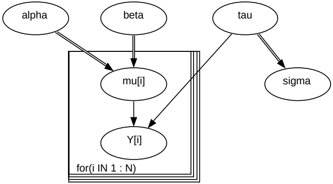
$$Y_i \sim N(\mu_i, \sigma^2) \text{ with } \mu_i = \alpha + \beta x_i \text{ for } i = 1, \dots, N.$$

we must also specify prior distributions for α, β, σ:





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)
    }
```

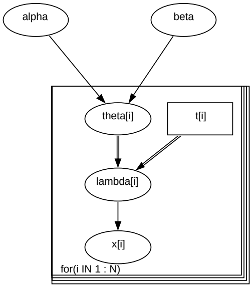
Data and BUGS model for pumps

The number of failures X_i is assumed to follow a Poisson distribution with parameter $\theta_i t_i, i = 1, \ldots, 10$
where θ_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.

Pump	1	2	3	4	5	6	7	8	9	10
t_i	94.5	15.7	62.9	126	5.24	31.4	1.05	1.05	2.01	10.5
x_i	5	1	5	14	3	19	1	1	4	22

A gamma prior distribution is adopted for the failure rates:
$\theta_i \sim \Gamma(\alpha, \beta), i = 1, \ldots, 10$

Gamma model for pumpdata



Failure of 10 power plant pumps.

BUGS program for pumps

With suitable priors the program becomes

```
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)
    }
```

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:

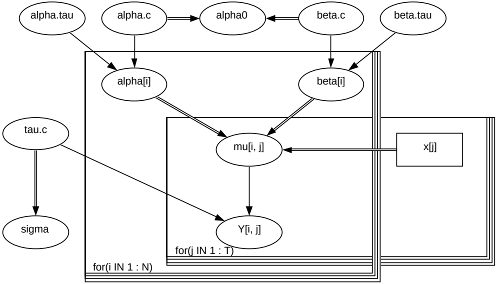
$$Y_{ij} \sim \mathcal{N}(\alpha_i + \beta_i(x_j - \bar{x}), \tau_c^{-1})$$

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 τ represents the precision (inverse variance) of a normal distribution. Interest particularly focuses on the intercept at zero time (birth), denoted $\alpha_0 = \alpha_c - \beta_c \bar{x}$.

Growth of rats



Growth of 30 young rats.

When exact computation is infeasible, Markov chain Monte Carlo (MCMC) methods are used.

An MCMC method for the *target distribution* π^* on $\mathcal{X} = \mathcal{X}_V$ constructs a Markov chain $X^0, X^1, \ldots, X^k, \ldots$ with π^* as *equilibrium distribution.*

For the method to be useful, π^* 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 χ_A is the indicator function of the set A.

Suppose we have sampled $X^1 = x^1, \ldots, X^n = x^{k-1}$ and we next wish to sample X^k. We choose a *proposal kernel* $g_k(y \mid z)$ and proceed as:

1. `Draw` $y \sim g_k(\cdot \mid x^{k-1})$. `Draw` $u \sim U(0,1)$.
2. `Calculate acceptance probability`

$$\alpha = \min\left\{1, \frac{\pi^*(y) g_k(x^{k-1} \mid y)}{\pi^*(x^{k-1}) g_k(y \mid x^{k-1})}\right\} \quad (1)$$

3. `If` $u < \alpha$ `set` $x^k = y$; `else set` $x^k = x^{k-1}$.

The samples $x^1, \ldots, x^M$ generated this way will form an ergodic Markov chain that, under certain conditions, has $\pi^*(x)$ as its stationary distribution.

A particularly simple special case is the *single site Gibbs sampler* where the update distributions all have the form of so-called *full conditional distributions*

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

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

4. Continue similarly to update x^k to x^{k+1} and so on.

The Gibbs sampler is just the Metropolis–Hastings algorithm with full conditionals as proposals.

For then the acceptance probabilities in (1) become

$$\begin{aligned}\alpha &= \min\left\{1, \frac{\pi^*(y_i \mid x_{V\setminus i}^{k-1})\pi^*(x^{k-1})}{\pi^*(x_i^{k-1} \mid x_{V\setminus i}^{k-1})\pi^*(y_i, x_{V\setminus i}^{k-1})}\right\} \\ &= \min\left\{1, \frac{\pi^*(y_i, x_{V\setminus i}^{k-1})\pi^*(x^{k-1})}{\pi^*(x_i^{k-1}, x_{V\setminus i}^{k-1})\pi^*(y_i, x_{V\setminus i}^{k-1})}\right\} = 1.\end{aligned}$$

Properties of Gibbs sampler

With positive joint target density $\pi^*(x) > 0$, the Gibbs sampler is ergodic with π^* as the unique equilibrium.

In this case the distribution of X^n converges to π^* 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.

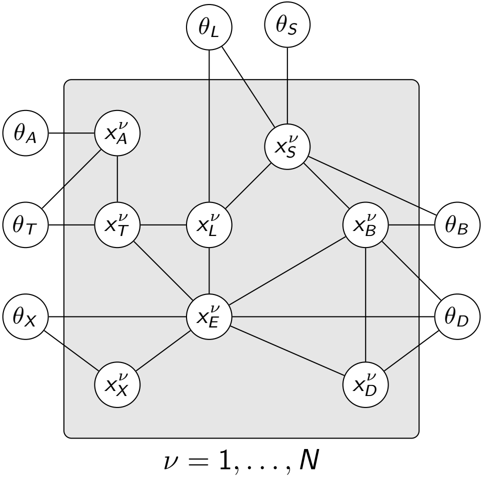
For a directed graphical model, the density of full conditional distributions are:

$$\begin{aligned} f(x_i \mid x_{V\setminus i}) &\propto \prod_{v \in V} f(x_v \mid x_{\mathrm{pa}(v)}) \\ &\propto f(x_i \mid x_{\mathrm{pa}(i)}) \prod_{v \in \mathrm{ch}(i)} f(x_v \mid x_{\mathrm{pa}(v)}) \\ &= f(x_i \mid x_{\mathrm{bl}(i)}), \end{aligned}$$

x where $\mathrm{bl}(i)$ is the *Markov blanket* of node i:

$$\mathrm{bl}(i) = \mathrm{pa}(i) \cup \mathrm{ch}(i) \cup \left\{ \cup_{v \in \mathrm{ch}(i)} \mathrm{pa}(v) \setminus \{i\} \right\}.$$

Note that *the Markov blanket is just the neighbours of i in the moral graph*: $\mathrm{bl}(i) = \mathrm{ne}^m(i)$.



Moral graph of chest clinic example.

There are many ways of sampling from a density f which is *known up to normalization*, i.e. $f(x) \propto h(x)$.

For example, one can use 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.