BCCS 2008/09: Graphical models and complex stochastic systems: Take-home open-book test sketch solutions and comments

1. Advantages of thinking graphically:

   In model formulation: complex models formed out of simple local components, which are easier to think about, rigorously linked mathematically; assists verbal explanation and interpretation; encourages parsimony (preferring simple explanations to complicated ones); supports (Bayesian) hierarchical models.

   In computation: simplifies computation of full conditionals when constructing MCMC methods like the Gibbs sampler; enables implementaion of junction-tree algorithms (probability propagation); assists in algorithmic organisation in many other kinds of computation; allows GUI representation for input and output of models.

2. (a) The likelihood is $L(\theta) = \prod_{i=1}^{n} p(x_i|\theta)$, which can be simplified to $\exp(-\theta \sum_i t_i)\theta^{\sum_i x_i} \prod_i [t_i^{x_i}/x_i!]$. Thus the log-likelihood is $\ell(\theta) = -\theta \sum_i t_i + \log(\theta) \sum_i x_i$, plus terms not involving θ, which do not affect the next step.

   Differentiating with respect to θ (the log-likelihood is clearly differentiable for $\theta > 0$), we find $\ell'(\theta) = -\sum_i t_i + (1/\theta) \sum_i x_i$, which is zero for $\theta = \sum_i x_i / \sum_i t_i$. To check that this does give a maximum, we could either differentiate again, or simply note by inspection that $\ell'(\theta)$ is a decreasing function of θ. Thus the maximum likelihood estimate of θ is $\sum_i x_i / \sum_i t_i$

   (b) The posterior distribution is proportional to the product of the likelihood and the prior, that is, to

$$\exp(-\theta \sum_i t_i)\theta^{\sum_i x_i} \prod_i [t_i^{x_i}/x_i!] \times \beta^\alpha \theta^{\alpha-1} \exp(-\beta\theta)/\Gamma(\alpha)$$

   for $\theta > 0$, 0 otherwise. Since we only care about proportionality (in θ), we can simplify to see

$$p(\theta|x_1, x_2, \ldots, x_n) \propto \exp(-\theta \sum_i t_i)\theta^{\sum_i x_i}\theta^{\alpha-1} \exp(-\beta\theta) = \exp(-\theta[\sum_i t_i + \beta])\theta^{\sum_i x_i + \alpha},$$

   which, by comparison with the gamma density in the question, we can recognise as the Gamma$(\sum_i x_i + \alpha, \sum_i t_i + \beta)$ distribution.

   So the posterior expectation is the ratio of the parameters, that is $(\sum_i x_i + \alpha)/(\sum_i t_i + \beta)$. Note that if *either* $\alpha = \beta = 0$, or if α and β are negligible compared with $\sum_i x_i$ and $\sum_i t_i$ (which will apply as $n \to \infty$), then the posterior expectation equals the maximum likelihood estimator.

3. (a) Using the notes, $r_1(x_1) = \sum_{x_0} g_1(x_0, x_1) r_0(x_0) = \sum_{x_0} p(x_0)p(x_1|x_0)p(y_1|x_1) \times 1 = \sum_{x_0} p(x_0, x_1, y_1) = p(x_1, y_1)$. Now proceed by induction: if it is true that $r_t(x_t) = p(x_t, y_{\le t})$, then $r_{t+1}(x_{t+1}) = \sum_{x_t} g_{t+1}(x_t, x_{t+1}) r_t(x_t)$
   $= \sum_{x_t} p(x_{t+1}|x_t)p(y_{t+1}|x_{t+1})p(x_t, y_{\le t}) = \sum_{x_t} p(x_t, x_{t+1}, y_{\le t}, y_{t+1})$ (see (*) below)
   $= \sum_{x_t} p(x_t, x_{t+1}, y_{\le t+1}) = p(x_{t+1}, y_{\le t+1})$. Thus it is true for all t.
   Finally, $p(x_t|y_{\le t}) = p(x_t, y_{\le t})/p(y_{\le t}) = p(x_t, y_{\le t})/\sum_{x_t} p(x_t, y_{\le t})$, so substitute and we are done.

(*) To see why $p(x_{t+1}|x_t)p(y_{t+1}|x_{t+1})p(x_t, y_{\leq t}) = p(x_t, x_{t+1}, y_{\leq t}, y_{t+1})$, note that it is always true that

$$p(x_t, x_{t+1}, y_{\leq t}, y_{t+1}) = p(x_t, y_{\leq t})p(x_{t+1}|x_t, y_{\leq t})p(y_{t+1}|x_{t+1}, x_t, y_{\leq t}).$$

But $p(x_{t+1}|x_t, y_{\leq t}) = p(x_{t+1}|x_t)$ since $x_{t+1} \perp\!\!\!\perp y_{\leq t} \mid x_t$, and $p(y_{t+1}|x_{t+1}, x_t, y_{\leq t}) = p(y_{t+1}|x_{t+1})$ since $y_{t+1} \perp\!\!\!\perp x_t, y_{\leq t} \mid x_{t+1}$.

(b) $p(x_t, y_{\leq t+1})$ is equal to $\sum_{x_{t+1}} p(x_t, x_{t+1}, y_{\leq t}, y_{t+1})$. Then using (*) above again, we see that the probability can be factorised as indicated.

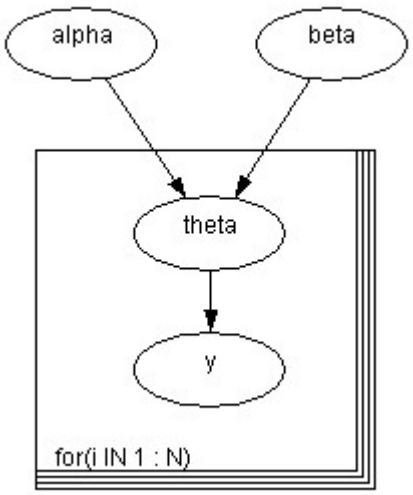
An efficient algorithm is therefore to compute all the $r_t(x_t)$ functions by forward recursion, and then to use the identity above to find $p(x_t, y_{\leq t+1})$, and finally to normalise to get the required probabilities $p(x_t|y_{\leq t+1}) = p(x_t, y_{\leq t+1}) / \sum_{x_t} p(x_t, y_{\leq t+1})$.

(c)

$$p(x_{t+1}|y_{\leq t}) = \sum_{x_t} p(x_{t+1}, x_t|y_{\leq t}) = \sum_{x_t} p(x_t|y_{\leq t})p(x_{t+1}|x_t, y_{\leq t}) = \sum_{x_t} p(x_t|y_{\leq t})p(x_{t+1}|x_t)$$

since $x_{t+1} \perp\!\!\!\perp y_{\leq t} \mid x_t$. Similarly to (b), an efficient algorithm is therefore to compute all the $r_t(x_t)$ functions by forward recursion, use (a) to find $p(x_t|y_{\leq t})$ and then to use the identity above to find $p(x_{t+1}|y_{\leq t})$.

4. (a) The DAG looks like this:



(b)

$$p(\alpha, \beta, \{\theta_i\}, \{y_i\}) = e^{-\alpha}e^{-\beta} \prod_{i=1}^{n} \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_i^{\alpha-1}(1-\theta_i)^{\beta-1} \frac{(y_i+r-1)!}{y_i!(r-1)!} \theta_i^r (1-\theta_i)^{y_i} \right)$$

(c)

$$p(\theta_i| \ldots) \propto p(\alpha, \beta, \{\theta_i\}, \{y_i\}) \propto \theta_i^{\alpha-1}(1-\theta_i)^{\beta-1}\theta_i^r(1-\theta_i)^{y_i} = \theta_i^{\alpha+r-1}(1-\theta_i)^{\beta+y_i-1}$$

which we recognise as a Beta distribution, in fact Beta$(\alpha + r, \beta + y_i)$. This is a standard distribution, so Gibbs sampling is straightforward for θ_i.

(d)

$$p(\alpha|\ldots) \propto p(\alpha, \beta, \{\theta_i\}, \{y_i\}) \propto e^{-\alpha} \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)^n \left(\prod \theta_i\right)^{\alpha-1}$$

which is not a standard distribution, so Gibbs sampling would be problematical for α, and similarly for β. But we can use another standard MCMC method instead, for example Metropolis-Hastings sampling.

(e) Bugs code for this model looks like this (as in the NegBin.zip file, online):

```
model
{
for( i in 1 : N ) {
y[i] ~ dnegbin(theta[i],2)
theta[i] ~ dbeta(alpha,beta)
}
alpha ~ dexp(1)
beta ~ dexp(1)
}
```

If you run this, following the instructions in the readme file, you obtain various statistics and graphs from the run, summarising both the performance of the simulation, and the posterior distributions that it computes. Briefly, in performance terms, everything seems fine, the run stabilises quickly, and autocorrelations in the simulated series are low (less low for α and β than the other parameters, but still quite good).

The posterior means for the θ_i range from 0.126 (θ_{10}) to 0.575 (θ_6), with posterior standard deviations typically about half the size of the mean or a little less. The posterior distributions for θ_i are rather symmetric when the posterior mean is large, skewed to the right when it is small. The hyperparameters α and β are estimated to be about 1.21 and 2.63, and both have posterior distributions skewed to the right.

5. To answer these questions, we need to check the global Markov property for the directed graph, as in section 3.3 of the notes. That is, we start with the original graph, then

(i) delete all nodes that are not in A, B or C, or *ancestors* of nodes in A, B or C, and all arrows into or out of deleted nodes.

(ii) add an edge between each pair of parents that are not already connected (*moralisation*).

(iii) drop directions on all the arrows

Then we look at the resulting graph to see if C *separates* A and B – that is, if it is impossible to find a path from a node in A to a node in B that does not pass through C. The global Markov property says that if C separates A and B, then $A \perp\!\!\!\perp B \mid C$.

The results are

(a) Yes (which we can see also more directly from the local Markov property, since (b, c) are the parents of e).

(b) No. There is a path $b \to a \to c \to f \to h$ in the resulting graph. So you *cannot* infer that $b \perp\!\!\!\perp h \mid d$. Note that this is not the same as concluding that b and h are *dependent* given d, it is just a question that cannot be settled given the information stated.

(c) Yes, by the global Markov property.

(d) Yes, by the global Markov property. Also, to see this algebraically, note that

$$\begin{aligned} p(a,c,d,e,g,h) &= \sum_{b,f} p(a,b,c,d,e,f,g,h) \\ &= \sum_{b,f} p(a)p(b|a)p(c|a)p(d|a)p(e|b,c)p(f|c,d)p(g|e,f)p(h|f) \\ &= (\sum_{b} p(a)p(b|a)p(c|a)p(d|a)p(e|b,c)) \times (\sum_{f} p(f|c,d)p(g|e,f)p(h|f)) \end{aligned}$$

which shows that it is a function of (a,c,d,e) alone, times a function of (c,d,e,g,h) alone, which in turn proves that $a \perp\!\!\!\perp (g,h) \mid (c,d,e)$ as in, for example, question 1 of exercise sheet 3.

This proof is really a special case of the one for the general case, presented in Section 3.3 of the notes.