T-61.5040 Oppivat mallit ja menetelmät
T-61.5040 Learning Models and Methods
Pajunen, Viitaniemi

Solutions to exercise 8, 16.3.2007

Problem 1.

Assume that the posterior $p(\theta|y)$ is an N-dimensional Normal $N(\theta|0, \sigma^2 I)$, and for $g(\theta)$ we use another Normal $N(\theta|0, \sigma_g^2 I)$ to do rejection sampling on the posterior.

First, let us look at the variances of $p(\theta|y)$ and $g(\theta)$. There must be a known constant M for which

$$\begin{aligned} M \geq \frac{p(\theta|y)}{g(\theta)} &= \frac{\sigma^{-N} \exp(-\frac{1}{2}\theta^T(\sigma^2 I)^{-1}\theta)}{\sigma_g^{-N} \exp(-\frac{1}{2}\theta^T(\sigma_g^2 I)^{-1}\theta)} \\ &= (\frac{\sigma_g}{\sigma})^N \exp(-\frac{1}{2}\theta^T\theta(\sigma^{-2} - \sigma_g^{-2})) \end{aligned}$$

Now the argument of exp must be negative so that the quotient of $p(\theta|y)$ and $g(\theta)$ does not grow without bounds. This gives $\sigma^{-2} - \sigma_g^{-2} > 0$ and thus $\sigma^2 < \sigma_g^2$; the variance of the sampling distribution must be larger than the posterior variance. We choose $\sigma_g = (1+\epsilon)\sigma$ where $\epsilon > 0$.

Now $M \geq \frac{p(\theta|y)}{g(\theta)}$ should hold for all θ. Especially when $\theta = 0$, the $\exp(\ldots)$ term is largest (now that we have chosen $\sigma_g^2 > \sigma^2$), namely then $\exp(\ldots) = 1$. In this case

$$M = (\frac{\sigma_g}{\sigma})^N = \exp(N \log \frac{\sigma_g}{\sigma}) = \exp(N \log(1+\epsilon)) \approx exp(N\epsilon) \text{ when } \epsilon \text{ is small}$$

For example, if $N = 1000$ and $\epsilon = 0.1$ we obtain $M \approx 2.5 \cdot 10^{41}$. A sample θ is accepted with probability $\frac{p(\theta|y)}{Mg(\theta)}$, so in this case about one in 10^{41} samples will be accepted. We see that rejection sampling is difficult especially in high dimensions. On the other hand, rejection sampling is the only simple method for obtaining samples directly from $p(\theta|y)$.

Problem 2.

Since we assume that there is a unique stationary distribution for the Markov chain, it is enough to show that the posterior $p(\theta|y)$ is stationary. Assume that θ^n is from the posterior distribution. Choose any values θ_1 and θ_2 such that

$$p(\theta_1|y) \geq p(\theta_2|y). \tag{1}$$

First we compute the probability that the simulation is at θ_2 at time n and at θ_1 at time $n+1$. This is

$$\begin{aligned} P_{21} &= p(\theta^n = \theta_2, \theta^{n+1} = \theta_1|y) \\ &= p(\theta^{n+1} = \theta_1|\theta^n = \theta_2, y)p(\theta^n = \theta_2|y). \end{aligned}$$

The first probability is the transition probability from θ_2 to θ_1: this is $J(\theta_1|\theta_2)p_r$. The second probability is by assumption $p(\theta_2|y)$. Since r is at least one by (1) and thus $p_r = 1$, the transition probability will be

$$P_{21} = p(\theta_2|y)J(\theta_1|\theta_2).$$

The probability

$$P_{12} = p(\theta^n = \theta_1, \theta^{n+1} = \theta_2|y)$$

can be obtained as above, but now r is at most one (and thus $p_r = r$):

$$P_{12} = p(\theta_1|y)J(\theta_2|\theta_1)r.$$

Substituting $r = p(\theta_2|y)/p(\theta_1|y)$ we obtain

$$P_{12} = p(\theta_2|y)J(\theta_2|\theta_1).$$

Since J is symmetric, we get $P_{12} = P_{21}$. This means that the distribution $p(\theta^n, \theta^{n+1}|y)$ is symmetric w.r.t. θ^n and θ^{n+1}. We know that $p(\theta^n|y) = p(\theta|y)$. Then

$$\begin{aligned} p(\theta^{n+1}|y) &= \int p(\theta^n, \theta^{n+1}|y)d\theta^n \\ &= \int p(\theta^n, \theta^{n+1}|y)d\theta^{n+1} \\ &= p(\theta^n|y) = p(\theta|y). \end{aligned}$$

This means that if the simulation has the posterior distribution at time n, then it will have it at time $n+1$, showing that it is stationary.

If the posterior has two or more separate areas as in the problem statement (p_1 and p_2), then it is possible that there is more than one stationary distribution for the Markov chain. For example, if the jumping distribution prevents jumps of distance 1 or more, it is impossible to jump from p_1 to p_2, and vice versa. This is avoided if the jumping distribution can jump to any point with positive probability. Normal distribution is one such jumping distribution.

Comments: With quite general assumptions, mainly that the simulation has a positive probability of reaching any point θ, one can show that the simulation actually converges to the stationary distribution. In practice, by choosing a suitable jumping distribution the assumptions are fulfilled.

Problem 3.

i) First compute $p(\mu|\sigma, y)$. This is again the posterior for inferring the mean of a Normal distribution when the variance is known. The results obtained earlier give

$$p(\mu|\sigma, y) = N(\mu|\frac{\sigma_0^{-2}\mu_0 + \sigma^{-2}\sum_i y_i}{\sigma_0^{-2} + n\sigma^{-2}}, (\sigma_0^{-2} + n\sigma^{-2})^{-1})$$

The distribution $p(\sigma^2|\mu, y)$ is the posterior for inferring the unknown Normal variance when the mean μ is known. Since the prior $p(\sigma^2)$ is Inverse-Gamma, then the posterior is

$$p(\sigma^2|\mu, y) = IG(\sigma^2|\frac{n}{2} + a, \frac{1}{2}(2b + ny))$$

ii) The posterior mean of μ is

$$\mathrm{E}(\mu|y) = \int \mu p(\mu|y) d\mu \quad (*)$$

Since the simulated values obtained via Gibbs sampling approximate the full posterior $p(\mu, \sigma|y)$, it is not necessarily trivial how the expectation $(*)$ is approximated. If we had samples $z_1, \ldots, z_N$ from $p(\mu|y)$, then (*) could be Monte Carlo-approximated as

$$\mathrm{E}(\mu|y) \approx \frac{1}{N} \sum_i z_i$$

Write $(*)$ using the full posterior as

$$\begin{aligned} \mathrm{E}(\mu|y) &= \int \mu p(\mu|y) d\mu \\ &= \int \mu \int p(\mu, \sigma^2|y) d\sigma^2 d\mu \\ &= \int \int \mu p(\mu, \sigma^2|y) d\sigma^2 d\mu \end{aligned}$$

Now we have an integral over the full posterior. If we choose the function to be integrated over as $h(\mu, \sigma^2) = \mu$, then we obtain the posterior mean $\mathrm{E}(\mu|y)$ as an integral over the full posterior. But Monte Carlo approximation is now

$$\mathrm{E}(\mu|y) \approx \frac{1}{N} \sum_i h(\mu_i, \sigma_i^2) = \frac{1}{N} \sum_i \mu_i$$

This holds in general, meaning that *marginalization* is trivial when using simulated posteriors $\theta^1, \ldots, \theta^N$: just ignore the components of θ^i you are not interested in.

Problem 4.

i) The posterior is zero when any y_i is less than a. Therefore as one of the two scalar functions determining the posterior we must use is $y^* = \min y_i$. This gives the posterior as

$$p(a, b|y) \propto p(y|a, b)p(a, b) = \begin{cases} 0, & y^* < a \\ b^{-1} \prod_{i=1}^n b \exp(-b(y_i - a)), & y^* \geq a \end{cases}$$

The case $y^* \geq$ a can be written as

$$b^{n-1} \exp(-b \sum_i y_i) \exp(abn)$$

so the second scalar function is $\sum_i y_i$. In statistics, y^* and $\sum_i y_i$ are called *sufficient statistics* because they summarise all the information data contains about the unknown quantities.

ii) First we compute $p(b|a, y)$. Using Bayes' Theorem we get

$$\begin{aligned} p(b|a, y) &\propto p(y|a, b)p(b|a) \\ &= b^{-1} \prod_i b \exp(-b(y_i - a)), \text{ when } y^* \geq a \\ &= b^{n-1} \exp(-\sum_i (y_i - a)b) \\ &= Gamma(b|n, \sum_i (y_i - a)) \end{aligned}$$

This can be simulated when $\sum_i y_i$ and a are known.

Then compute $p(a|b, y)$. Similarly to above, this would be

$$p(a|b, y) \propto p(y|a, b)p(a|b) = p(y|a, b)p(a, b)/p(b).$$

Suppose this was difficult to simulate. We can use the fact that we don't have to directly simulate a: we could as well simulate a function of a and b, provided that we can solve a from it. Write $z = \exp(abn)$. Then

$$p(z|b, y) \propto p(y|z, b)p(z|b) = p(y|a, b)p(z|b) = b^n \exp(-\sum_i (y_i - a)b)p(z|b) = b^n \exp(-b \sum_i y_i) z p(z|b)$$

To compute $p(z|b) = p(\exp(abn)|b)$ we need to use the formula for transforming variables: $p(z) = p(a) \left|\frac{da}{dz}\right|$ Since $a = (bn)^{-1} \log z$, we get $da/dz = (bnz)^{-1}$. Also, $p(a) \propto 1$ so we get

$$p(z|b) = p(z) \propto z^{-1}$$

Then

$$p(z|b, y) \propto b^n \exp(-b \sum_i y_i)$$

which is independent of z and a. Therefore z has a uniform distribution. It remains to find the interval where this distribution is uniform. First, since $a > 0$, it follows that $z > 1$. The likelihood is zero when $a > y^*$ so it must hold that $z = \exp(abn) \leq \exp(y^* bn)$. Therefore $p(z|b, y) = U(1, \exp(y^* bn))$.

Now the Gibbs Sampler is ready:
1. Choose initial values a_0 and b_0.
2. Simulate b_1 from the Gamma distribution using a_0
3. Simulate $z_1 = \exp(a_1 b_1 n)$ from the uniform distribution using b_1.
4. Solve a_1 by $(b_1 n)^{-1} \log z_1$.
5. Iterate by going back to 2.