Petter Mostad
Applied Mathematics and Statistics
Chalmers

**Suggested solutions for**
**MSA101 / MVE187 Computational methods for Bayesian statistics**
**Exam 2 January 2018**

1. (a) One may use rejection sampling: In fact one may use the uniform distribution on $[0, 3]$ as a proposal distribution. The algorithm is then:

   - Generate uniform variables $U_1$ and $U_2$ on the interval $[0, 1]$.
   - If $\pi(3U_1) < BU_2$ then store the sampled value $x = 3U_1$, otherwise return to the first point.

   (b) As the area of the rectangle with baseline $[0, 3]$ and height $B$ is $3B$ and the area under the density $\pi$ is 1, the probability of not rejecting the proposed value for $x$ is $1/3B$. This shows directly how the algorithm can be inefficient if $B$ is large.

2. We get

$$\begin{aligned}
\pi(x) &= \int_0^\infty \pi(x \mid y)\pi(y)\, dy \\
&= \int_0^\infty y \exp(-yx) \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} \exp(-\beta y)\, dy \\
&= \frac{\beta^\alpha}{\Gamma(\alpha} \int_0^\infty y^\alpha \exp(-(x+\beta)y)\, dy \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot \frac{\Gamma(\alpha+1)}{(\beta+x)^{\alpha+1}} \\
&= \frac{\alpha}{\beta}\left(\frac{\beta+x}{\beta}\right)^{-(\alpha+1)} = \frac{\alpha}{\beta}\left(1+\frac{x}{\beta}\right)^{-(\alpha+1)}.
\end{aligned}$$

3. As $g_i(\theta \mid x)$ is the posterior corresponding to the prior $f_i(\theta)$, we get the corresponding prior predictive distribution $h_i(x)$ defined with

$$h_i(x) = \frac{\pi(x \mid \theta) f_i(\theta)}{g_i(\theta \mid x)}$$

where the $\theta$ will disappear from the expression on the right. Thus

$$\begin{aligned}
\pi(\theta \mid x) &= \frac{\pi(x \mid \theta)\pi(\theta)}{\pi(x)} \\
&= \frac{\sum_{i=1}^{k} \pi(x \mid \theta) c_i f_i(\theta)}{\int \sum_{i=1}^{k} \pi(x \mid \theta) c_i f_i(\theta)\, d\theta} \\
&= \frac{\sum_{i=1}^{k} c_i h_i(x) g_i(\theta \mid x)}{\int \sum_{i=1}^{k} c_i h_i(x) g_i(\theta \mid x)\, d\theta} \\
&= \frac{\sum_{i=1}^{k} c_i h_i(x) g_i(\theta \mid x)}{\sum_{i=1}^{k} c_i h_i(x)} \\
&= \sum_{i=1}^{k} \left[ \frac{c_i h_i(x)}{\sum_{j=1}^{k} c_j h_j(x)} \right] g_x(\theta \mid x)
\end{aligned}$$

4. The difference lies in the interpretation. In a causal network, there is a possibility to "set" or "fix" variables, which is a different action from just observing it. When a variable is set, the conditonal distribution of the remaining variables correspond to the conditional distribution in the Bayesian network where the incoming edges to the set variable have been removed. If the network is not a causal network but only a Bayesian network, there is only a possibliity to observe variables, not to set or fix them.

5. (a) First, an initial reasonable value for $x$ is simulated; call it $x_0$. Then, for $i = 1, \ldots, n$,
   - Simulate $y$ using the proposal density $q(y \mid x_{i-1})$.
   - Compute the acceptance probability

$$p = \min\left(1, \frac{f(y)q(x_{i-1} \mid y)}{f(x_{i-1})q(y \mid x_{i-1})}\right)$$

   - Set $x_i = y$ with probability $p$, otherwise, set $x_i = x_{i-1}$.

   (b) Let $K(x, y)$ denote the probability (density) that the chain has value $y$ given that it has value $x$ as the previous step. The detailed balance condition for a density $f$ is then that we have, for all $x$ and $y$,

$$f(x)K(x, y) = f(y)K(y, x)$$

   If a Markov chain satisfies the detailed balance condition for a density $f$, it is fairly easy to see that $f$ is a stationary distribution for the chain. If the Markov chain is defined by the Metropolis Hastings algorithm above, one can show that it satisfies the detailed balance condition for the density $f$. Thus, if it has a unique stationary distribution, it must be $f$.

   (c) The general idea may be explained as follows: For each chain, remove the initially simulated values (the "burn-in") and make a "thinning" by selecting only each $k$'th

of the remaining values, for some $k$, so that one believes the values remaining after this are an approximate random sample from the density. As a check of whether this is true, one may compare the variability (e.g., the variance) within the values from each chain to the variability within the set of all the simulated values. If the latter is clearly greater, it indicates that the values from each chain to some extent depend on the starting value of the chain, and that convergence has not been reached.

6. (a) A Gibbs sampler would anternate between simulating $\mu$ given $\tau_1$ and simulating $\tau_1$ given $\mu$, starting with some value for, e.g., $\tau_1$. When fixing $\tau_1$ the posterior $\pi(\mu \mid x, \tau_1)$ can be found using conjugacy:

$$\begin{aligned}
\pi(\mu \mid x, \tau_1) &\propto_\mu \pi(\mu)\pi(x \mid \mu, \tau_1) \\
&\propto_\mu \exp\left(-\frac{\tau_0}{2}(\mu-\mu_0)^2\right)\exp\left(-\frac{\tau_1}{2}(x-\mu)^2\right) \\
&= \exp\left(-\frac{1}{2}\left(\tau_0\mu^2 - 2\tau_0\mu\mu_0 + \tau_0\mu_0^2 + \tau_1\mu^2 - 2\tau_1\mu x + \tau_1 x^2\right)\right) \\
&\propto_\mu \exp\left(-\frac{1}{2}\left((\tau_0+\tau_1)\mu^2 - 2(\tau_0\mu_0 + \tau_1 x)\mu\right)\right) \\
&\propto_\mu \exp\left(-\frac{1}{2}(\tau_0+\tau_1)\left(\mu - \frac{\tau_0\mu_0 + \tau_1 x}{\tau_0+\tau_1}\right)^2\right)
\end{aligned}$$

so that

$$\mu \mid x, \tau_1 \sim \text{Normal}\left(\frac{\mu_0\tau_0 + x\tau_1}{\tau_0+\tau_1}, \frac{1}{\tau_0+\tau_1}\right)$$

When fixing $\mu$ the posterior $\pi(\tau_1 \mid x, \mu)$ can also be found using conjugacy:

$$\begin{aligned}
\pi(\tau_1 \mid x, \mu) &\propto_{\tau_1} \pi(\tau_1)\pi(x \mid \mu, \tau_1) \\
&\propto_{\tau_1} \tau_1^{\alpha-1}\exp(-\tau_1\beta)\cdot\tau_1^{1/2}\exp\left(-\frac{\tau_1}{2}(x-\mu)^2\right) \\
&= \tau_1^{\alpha-1/2}\exp\left(-\tau_1\left(\beta + \frac{1}{2}(x-\mu)^2\right)\right)
\end{aligned}$$

so that

$$\tau_1 \mid x, \mu \sim \text{Gamma}\left(\alpha + \frac{1}{2}, \beta + \frac{1}{2}(x-\mu)^2\right)$$

(b) The algorithm to use is the EM algorith, with $\mu$ as the "augmented data". The full log posterior becomes (up to constants $C_1$ and $C_2$ not depending on $\tau_1$)

$$\begin{aligned}
&\log\left(\pi(x \mid \mu, \tau_1)\pi(\tau_1)\pi(\mu)\right) \\
= &\log\left[\frac{1}{\sqrt{2\pi/\tau_1}}\exp\left(-\frac{\tau_1}{2}(x-\mu)^2\right)\tau_1^{\alpha-1}\exp(-\beta\tau_1)\exp\left(-\frac{\tau_0}{2}(\mu-\mu_0)^2\right)\right] + C_1 \\
= &\frac{1}{2}\log\tau_1 - \frac{\tau_1}{2}(x-\mu)^2 + (\alpha-1)\log\tau_1 - \beta\tau_1 + C_2
\end{aligned}$$

For the E-step, we then need to consider the distribution of $\mu$ for fixed $x$ and fixed $\tau_1'$. We know from (a) that

$$\mu \mid x, \tau_1' \sim \text{Normal}(\mu_2, 1/\tau_2)$$

where $\mu_2 = \frac{\mu_0\tau_0 + x\tau_1'}{\tau_0 + \tau_1'}$ and $\tau_2 = \tau_0 + \tau_1'$. Thus we get that $E_{\tau_1'}(\mu) = \mu_2$ and

$$E_{\tau_1'}(\mu^2) = Var_{\tau_1'}(\mu) + E_{\tau_1'}(\mu)^2 = 1/\tau_2 + \mu_2^2$$

So for the M-step, we get, up to a constant $C_2$,

$$\begin{aligned} Q(\tau_1, \tau_1') &= E_{\tau_1'}\left[\log\left(\pi(x \mid \mu, \tau_1)\pi(\tau_1)\pi(\mu)\right)\right] \\ &= \left(\alpha - \frac{1}{2}\right)\log\tau_1 - \frac{\tau_1}{2}\left(x^2 - 2xE_{\tau_1'}(\mu) + E_{\tau_1'}(\mu^2)\right) - \beta\tau_1 + C_3 \\ &= \left(\alpha - \frac{1}{2}\right)\log\tau_1 - \frac{\tau_1}{2}\left(x^2 - 2x\mu_2 + \mu_2^2 + 1/\tau_2\right) - \beta\tau_1 + C_3 \end{aligned}$$

We find the value of $\tau_1$ maximizing this expression by differentiation, setting the result to zero:

$$\left(\alpha - \frac{1}{2}\right)\frac{1}{\tau_1} - \frac{1}{2}\left(2\beta + x^2 - 2x\mu_2 + \mu_2^2 + 1/\tau_2\right) = 0$$

The result is

$$\tau_1 = \frac{2\beta + x^2 - 2x\mu_2 + \mu_2^2 + 1/\tau_2}{2\alpha - 1}$$

In summary, the EM algorithm starts with a reasonable value for $\tau_1$. Then for each iteration, $\mu_2$ and $\tau_2$ are computed accoding to the formulas above, and then $\tau_1$ is computed according to the formula directly above.

7. The algorithm to use is the Viterbi algorithm. Roughly, it computes recursively the sequence $x_0, \ldots, x_i$ maximizing the probability of the data $y_0, \ldots, y_i$, and ending with specific values for $x_i$. For each such probability, the value of $x_{i-1}$ is also stored. When $i$ reaches $T$ the value of $x_T$ in the sequence maximizing the probaility can be found, and the previous values of the $x_i$ in the chain can be found by tracing back using the stored values of the $x_{i-1}$ for each $i$.

   In more detail, define, for $i = 0, \ldots, T$ and $j = 0, 1$,

   $$P(i, j) = \pi(y_0, \ldots, y_i, x_0, \ldots, x_i)$$

   where $x_i = j$ and the remaining $x_0, \ldots, x_{i-1}$ are such that the probability is maximized. Define also, for $i = 1, \ldots, T$ and $j = 0, 1$, $Q_{ij}$ as the value of $x_{i-1}$ in this sequence.

   These values can now be computed recursively. First,

   $$P(0, j) = \pi(y_0, x_0 = j) = \pi(y_0 \mid x_0 = j)\pi(x_0 = j).$$

To compute $P(i, j)$ and $Q_{ij}$ for $i > 0$, compute, for each possible value of $x_i$ and each possible value of $x_{i-1}$, $\pi(y_0, \ldots, y_i, x_0, \ldots, x_i)$ where $x_0, \ldots, x_{i-2}$ are such that the probability is maximized:

$$\begin{aligned}\pi(y_0, \ldots, y_i, x_0, \ldots, x_i) &= \pi(y_i \mid x_i)\pi(x_i \mid x_{i-1})\pi(y_0 \ldots, y_{i-1}, x_0, \ldots, x_{i-1}) \\ &= \pi(y_i \mid x_i)\pi(x_i \mid x_{i-1})P(i-1, x_{i-1}).\end{aligned}$$

For each $x_i$ find the $x_{i-1}$ maximizing the expression above, and set $Q_{ij}$ equal to this $x_{i-1}$. Then set $P(i, j)$ equal to the probability computed using this $x_{i-1}$.

In the end, the sequence $x_0, \ldots, x_T$ maximizing the probability for the data can be found by first finding the $j$ maximizing $P(T, j)$ and then tracing back the values of the $x_i$ using the computed $Q_{ij}$.

8. Sampling importance resampling is a way to obtain an approximate sample from a density $\pi(\theta)$ that is difficult to simulate from. One instead simulates a sample $\theta_1, \ldots, \theta_N$ from a proposal density $g(\theta)$ that is similar to $\pi(\theta)$ but easier to simulate from. Then, one resamples from this sample using probabilities

$$p_j = \frac{\pi(\theta_j)/g(\theta_j)}{\sum_{i=1}^{N} \pi(\theta_i)/g(\theta_i)}$$