Petter Mostad
Applied Mathematics and Statistics
Chalmers

Suggested solutions for
MSA101 / MVE187 Computational methods for Bayesian statistics
Exam 19 August 2019

1. (a) We have

$$\pi(p \mid x) \propto_p \pi(x \mid p)\pi(p) \propto_p p^x(1-p)^r p^{\alpha-1}(1-p)^{\beta-1} \propto_p p^{\alpha+x-1}(1-p)^{\beta+r-1}$$

so the posterior distribution becomes $\text{Beta}(\alpha+x, \beta+r)$.

(b) We may write

$$\begin{aligned}
\pi(x) &= \frac{\pi(\mid p)\pi(p)}{\pi(p \mid x)} \\
&= \frac{\frac{(x+r-1)!}{x!(r-1)!}p^x(1-p)^r\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1-p)^{\beta-1}}{\frac{\Gamma(\alpha+\beta+x+r)}{\Gamma(\alpha+x)\Gamma(\beta+r)}p^{\alpha+x-1}(1-p)^{\beta+r-1}} \\
&= \frac{\Gamma(x+r)\Gamma(\alpha+\beta)\Gamma(\alpha+x)\Gamma(\beta+r)}{\Gamma(x+1)\Gamma(r)\Gamma(\alpha+\beta+x+r)\Gamma(\alpha)\Gamma(\beta)}
\end{aligned}$$

(c) We get

$$\begin{aligned}
\pi(p \mid x) &\propto_p \pi(x \mid p)\pi(p) \\
&\propto_p p^x(1-p)^r\left[ap^{\alpha-1}(1-p)^{\beta-1} + bp^{\gamma-1}(1-p)^{\delta-1}\right] \\
&= ap^{\alpha+x-1}(1-p)^{\beta+r-1} + bp^{\gamma+x-1}(1-p)^{\delta+r-1}.
\end{aligned}$$

To find the exact form, it only remains to find the proportionality constant. Using some constant C, we rewrite as

$$\begin{aligned}
\pi(p \mid x) &= Cap^{\alpha+x-1}(1-p)^{\beta+r-1} + Cbp^{\gamma+x-1}(1-p)^{\delta+r-1} \\
&= Cap^{\alpha'-1}(1-p)^{\beta'-1} + Cbp^{\gamma'-1}(1-p)^{\delta'-1}
\end{aligned}$$

where we write $\alpha' = \alpha + x, \beta' = \beta + r, \gamma' = \gamma + x$, and $\delta' = \delta + r$. We can then write

$$\pi(p \mid x) = Ca\frac{\Gamma(\alpha'+\beta')}{\Gamma(\alpha')\Gamma(\beta')}\text{Beta}(p;\alpha',\beta') + Cb\frac{\Gamma(\gamma'+\delta')}{\Gamma(\gamma')\Gamma(\delta')}\text{Beta}(p;\gamma',\delta')$$

and integrating this over all possible p we get

$$Ca\frac{\Gamma(\alpha'+\beta')}{\Gamma(\alpha')\Gamma(\beta')} + Cb\frac{\Gamma(\gamma'+\delta')}{\Gamma(\gamma')\Gamma(\delta')} = 1$$

and thus

$$C = \frac{1}{a\frac{\Gamma(\alpha'+\beta')}{\Gamma(\alpha')\Gamma(\beta')} + b\frac{\Gamma(\gamma'+\delta')}{\Gamma(\gamma')\Gamma(\delta')}}.$$

From this we can write the exact posterior as

$$\pi(p \mid x) = \frac{a}{a\frac{\Gamma(\alpha'+\beta')}{\Gamma(\alpha')\Gamma(\beta')} + b\frac{\Gamma(\gamma'+\delta')}{\Gamma(\gamma')\Gamma(\delta')}} p^{\alpha'-1}(1-p)^{\beta'-1} + \frac{b}{a\frac{\Gamma(\alpha'+\beta')}{\Gamma(\alpha')\Gamma(\beta')} + b\frac{\Gamma(\gamma'+\delta')}{\Gamma(\gamma')\Gamma(\delta')}} p^{\gamma'-1}(1-p)^{\delta'-1}$$

or, if you like,

$$\pi(p \mid x) = \frac{a\frac{\Gamma(\alpha'+\beta')}{\Gamma(\alpha')\Gamma(\beta')}}{a\frac{\Gamma(\alpha'+\beta')}{\Gamma(\alpha')\Gamma(\beta')} + b\frac{\Gamma(\gamma'+\delta')}{\Gamma(\gamma')\Gamma(\delta')}} \text{Beta}(p;\alpha',\beta') + \frac{b\frac{\Gamma(\gamma'+\delta')}{\Gamma(\gamma')\Gamma(\delta')}}{a\frac{\Gamma(\alpha'+\beta')}{\Gamma(\alpha')\Gamma(\beta')} + b\frac{\Gamma(\gamma'+\delta')}{\Gamma(\gamma')\Gamma(\delta')}} \text{Beta}(p;\gamma',\delta').$$

2. (a) A possibility is rejection sampling, using an Exponential distribution with parameter 3 as a proposal density $q(x)$. We then get

$$\frac{p(x)}{q(x)} = \frac{Ce^{-3x}|\sin x|}{3\exp(-3x)} = \frac{1}{3}C|\sin x| \leq \frac{1}{3}C$$

so the quotient of densities is bounded by $C/3$. The rejection sampling algorithm then becomes:

i. Simulate $x \sim$ Exponential(3).
ii. Simulate $u \sim$ Uniform(0, 1).
iii. Accept x if $u\frac{C}{3}q(x) \leq p(x)$, i.e., if $u \leq |\sin x|$, otherwise we return to the first step.

It remains to describe how to simulate from the Exponential(3) distribution using only simulations from the standard uniform distribution: Note that the cumulative distribution function is $F(x) = 1 - \exp(-3x)$. Writing $u_0 = F(x)$, we get

$$x = F^{-1}(u_0) = -\frac{1}{3}\log(1-u_0)$$

Thus one can simulate $u_0 \sim$ Uniform(0, 1), and then compute $x = -\frac{1}{3}\log(1-u_0)$.

(b) In rejection sampling where $p(x) \leq Mq(x)$, the rejection rate is $1/M$. Thus in our case it is $3/C$, and an estimate for C is 3 divided by the frequency of rejection.

3. (a) We get

$$\begin{aligned} H[X] &= \text{E}\left[-\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(X-\mu)^2\right)\right)\right] \\ &= \text{E}\left[\frac{1}{2}\log(2\pi\sigma^2)0\frac{1}{2\sigma^2}(X-\mu)^2\right] \\ &= \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sigma^2 \\ &= \frac{1}{2} + \frac{1}{2}\log(2\pi(\sigma^2) \end{aligned}$$

(b)

$$\mathrm{KL}[p\|q] = \int p(x) \log \frac{p(x)}{q(x)}\, dx$$

4. In the answers formal arguments using d-separation can be given, but below, more intuitive arguments are given instead.

   (a) X and Y are independent: For example, one may remove, in this order, nodes Z_6, Z_5, Z_4, and Z_3, as they are not observed and nothing depend on them. This leaves a network where X and Y are in different components, and they are therefore necessarily independent.

   (b) X and Y are independent, with the same argument as above.

   (c) X and Y are dependent: The fixed value of Z_4 introduces a dependency between X and Z_3, and the fixed value of Z_5 introduces a dependency between Z_3 and Y. Thus there is a dependency between X and Y.

5. (a) We get

$$\begin{aligned}
& \log\left(\pi(\alpha, \beta, \gamma \mid y)\right) \\
= \; & C_0 + \log\left(\pi(\alpha)\pi(\beta)\pi(\lambda \mid \alpha, \beta)\pi(y \mid \lambda)\right) \\
= \; & C_1 + \log(\beta^2 \exp(-7\beta)) + \log\left(\frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} \exp(-\beta\lambda)\right) + \log(e^{-4\lambda}(4\lambda)^y) \\
= \; & C_2 + 2\log\beta - 7\beta + \alpha\log\beta - \log(\Gamma(\alpha)) + (\alpha - 1)\log\lambda - \beta\lambda - 4\lambda + y\log\lambda
\end{aligned}$$

   (b) Gibbs sampling would alternate between simulating from the three distributions $\pi(\alpha \mid \beta, \lambda, y)$, $\pi(\beta \mid \alpha, \lambda, y)$, and $\pi(\lambda \mid \alpha, \beta, y)$, first generating some starting values for the three parameters.

   From the above, we have that, up to a constant C_3 not depending on α,

$$\log(\pi(\alpha \mid \beta, \lambda, y)) = C_3 + \alpha\log\beta - \log(\Gamma(\alpha)) + (\alpha - 1)\log\lambda$$

   The easiest way to simulate from $\pi(\alpha \mid \beta, \lambda, y)$ would be to compute the above numbers for all 10 possible values of α, exponentiate these numbers, normalize them so that they sum to 1, and simulate from this discrete distribution.

   From (a) or directly from the specification of the model, one finds that the posterior $\pi(\beta \mid \alpha, \lambda, y)$ because of conjugacy is the Gamma distribution Gamma$(3 + \alpha, 7 + \lambda)$. Thus one may simulate from this distribution.

   From (a) or directly from the specification of the model, one finds that the posterior $\pi(\lambda \mid \alpha, \beta, y)$ because of conjugacy is the Gamma distribution Gamma$(\alpha + y, \beta + 4)$. Thus one may simulate from this distribution.

   (c) There is a large flexibility in the type of symmetric proposal function that could work, and in practice one would need to adapt the proposal function to optimize convergence speed. But as an example, one could for example change α with +1 or -1, with

a 50% chance for each choice, and independently add normally distributed variables with zero expectation and variance 1 (for example) to each of β and λ. The new value would be accepted with probability

$$a = \min\left(1, \frac{\pi(\alpha', \beta', \gamma' \mid y)}{\pi(\alpha, \beta, \gamma \mid y)}\right)$$

Note that the quotient in the expression above can be computed as the unknown factor is the same in both $\pi(\alpha', \beta', \gamma' \mid y)$ and $\pi(\alpha, \beta, \gamma \mid y)$. Using some starting point for the simulation, it would eventually converge to a sample from the posterior $\pi(\alpha, \beta, \lambda \mid y)$.

6. The goal of the EM algorithm would be to find the maximum posterior estimate $\hat{\theta}$ for the marginal posterior $\pi(\theta \mid y) \propto_\theta \pi(y \mid \theta)\pi(\theta)$.

   The algorithm is iterative and finds a sequence of parameters $\theta_0, \theta_1, \ldots$, so that each has a higher posterior density and the limit is a local maximum for the posterior. At each step, the algorithm can be formulated as going through two steps, the E step and the M step.

   In the E step one calculates the function

   $$f(\theta) = \mathrm{E}_z\left[\log(\pi(y \mid \theta, z)\pi(\theta, z))\right]$$

   where z has the density $\pi(z \mid y, \theta_{old})$, where θ_{old} is the value of θ in the previous iteration. In the M step, this function is mazimized to find the next value for θ.

7. Variational Bayes is a method which finds a density that approximates a target density $p(\theta)$ (it may be a posterior) for a parameter θ in cases where this density is difficult to find or simulate from. The idea is to find the density $q(\theta)$ in a more restricted class of densities which minimizes the Kullback-Leibler divergence (or distance) $\mathrm{KL}[q\|p]$. If for example θ consists of many components, $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$, then the restricted class could consist of products over densities for each of the components θ_i. One may show that a density q representing a local minumum for the KL divergence may be found with an iterative procedure that decreases the KL divergence at each step. Specifically, one may use an interative procedure that cycles through the components of θ, changing the density on one component at a time.

8. Assume given a Hidden Markov Model (HMM) with observed variables $Y_1, Y_2, \ldots, Y_m$ and an underlying Markov chain of hidden variables $X_1, X_2, \ldots, X_m$. Assume the X_i variables are discrete with a finite number of possible values. The goal of the Viterbi algorithm is to find a sequence $\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_m$ maximizing the probability of observing a given sequence $Y_1, \ldots, Y_m$. The algorithm passes two times through the sequence, first from 1 through m and then from m to 1. In the first pass, the values of X_i maximizing the likelihood, for each possible value of X_{i-1} is recorded. In the second pass, this information is used to construct the sought sequence.