Petter Mostad
Applied Mathematics and Statistics
Chalmers

Suggested solutions for
MSA100 / MVE186 Computer Intensive Statistical Methods
Re-exam 5 June 2017

1. The first statement is used within frequentist statistics. Its interpretation is as follows: There are statistics L_1 and L_2 defined in terms of a sample $x'_1, \ldots, x'_n$ from a normal distribution with expectation θ and variance 1, such that the stochastic interval $[L_1, L_2]$ contains θ with 95% probability; the values of these statistics computed on the given data is $L_1 = 2.3$ and $L_2 = 2.5$.

   The second statement is used within Bayesian statistics. Its interpretation is as follows: With some prior on θ (not specified in the question), the posterior probability that θ is in the interval $[2.3, 2.5]$ is 95%.

2. (a) If $p \sim \text{Beta}(\alpha, \beta)$ and $x \mid p \sim \text{Neg-Bin}(r, p)$, then

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

   Thus $p \mid x \sim \text{Beta}(\alpha + x, \beta + r)$, and the Beta family of distributions is conjugate to to the Negative Binomial distribution for the p parameter.

   (b) We get

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

   (c) The posterrior predictive is the same as the prior predictive using the posterior given x as the prior when predicting x_{NEW}. Using (a), we get in our case that

$$p \mid x \sim \text{Beta}(\alpha + x, \beta + r) = \text{Beta}(2+1, 1+2) = \text{Beta}(3, 3)$$

   and using (b) we then get

$$\begin{aligned}\pi(x_{NEW} \mid x) &= \binom{x_{NEW}+2-1}{x_{NEW}}\frac{\Gamma(3+3)\Gamma(3+x_{NEW})\Gamma(3+2)}{\Gamma(3)\Gamma(3)\Gamma(3+x_{NEW}+3+2)} \\ &= 720\frac{x_{NEW}+1}{(x_{NEW}+3)(x_{NEW}+4)(x_{NEW}+5)(x_{NEW}+6)(x_{NEW}+7)}\end{aligned}$$

(d) Let C be a varriable such that $C = 1$ means model 1 is used and $C = 2$ means model 2 is used. The Bayes factor B is equal to the likelihood ratio $\pi(x \mid C = 1)/\pi(x \mid C = 2)$. Thus it is equal to the ratio of the corresponding prior predictive distributions:

$$\begin{aligned} B &= \frac{\binom{x+r-1}{x} \frac{\Gamma(\alpha_1+\beta_1)\Gamma(\alpha_1+x)\Gamma(\beta_1+r)}{\Gamma(\alpha_1)\Gamma(\beta_1)\Gamma(\alpha_1+x+\beta_1+r)}}{\binom{x+r-1}{x} \frac{\Gamma(\alpha_2+\beta_2)\Gamma(\alpha_2+x)\Gamma(\beta_2+r)}{\Gamma(\alpha_2)\Gamma(\beta_2)\Gamma(\alpha_2+x+\beta_2+r)}} \\ &= \frac{\Gamma(\alpha_1+\beta_1)\Gamma(\alpha_1+x)\Gamma(\beta_1+r)\Gamma(\alpha_2)\Gamma(\beta_2)\Gamma(\alpha_2+x+\beta_2+r)}{\Gamma(\alpha_1)\Gamma(\beta_1)\Gamma(\alpha_1+x+\beta_1+r)\Gamma(\alpha_2+\beta_2)\Gamma(\alpha_2+x)\Gamma(\beta_2+r)} \end{aligned}$$

(e) We have, apriori, that $\pi(C = 1) = \pi(C = 2) = 0.5$. The posterior probabilities for the two models can be computed using Bayes formula on odds form, i.e.,

$$\frac{\pi(C = 1 \mid x)}{1 - \pi(C = 1 \mid x)} = B \cdot \frac{\pi(C = 1)}{\pi(C = 2)} = B \cdot 1$$

which solves to $\pi(C = 1 \mid x) = B/(1 + B)$. Using the posterior for each model computed in (a), the posterior probability density for p given x is

$$\frac{B}{1+B} \cdot \frac{\Gamma(\alpha_1 + x + \beta_1 + r)}{\Gamma(\alpha_1 + x)\Gamma(\beta_1 + r)} p^{\alpha_1+x-1}(1-p)^{\beta_1+r-1} + \frac{1}{1+B} \cdot \frac{\Gamma(\alpha_2 + x + \beta_2 + r)}{\Gamma(\alpha_2 + x)\Gamma(\beta_2 + r)} p^{\alpha_2+x-1}(1-p)^{\beta_2+r-1}$$

3. (a) The cumulative density for an Exponential distribution with parameter 2.7 is, for $x \geq 0$,

$$F(x) = 1 - \exp(-2.7x).$$

Writing $U = F(x)$, we get $x = -\log(1 - U)/2.7$. Thus, we may simulate from the distribution by first simulating U' uniformly on the interval $[0, 1]$, and then computing $x = -\log(U')/2.7$.

(b) Simulation may be done in several ways; one option is to simulate x from a Gamma distribution with parameters $\alpha = 2.7$ and $\beta = 9.1$ and output $1/x$. The Gamma distribution has density

$$\pi(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x)$$

and it is possible to simulate from it using rejection sampling. Differentiation shows that this density has its maximum at $(\alpha - 1)/\beta = 1.7/9.1 = 0.1868132$, where the density is then 2.652551. But we can also use the identity $\log(x) \leq x - 1$ to show that

$$\begin{aligned} \pi(x) &= \frac{\beta^\alpha}{\Gamma(\alpha)} \exp(-\beta x + (\alpha - 1)\log(x)) \leq \frac{\beta^\alpha}{\Gamma(\alpha)} \exp(-\beta x + (\alpha - 1)(x - 1)) \\ &= 45.94846 \cdot \exp(-7.4x) < 6.21 \cdot 7.4 \exp(-7.4x) \end{aligned}$$

when we use that $\alpha = 2.7$ and $\beta = 9.1$. Thus a simple solution is to use rejection sampling with an exponential distribution with parameter 7.4 as a proposal distribution, and $M = 6.21$. (More efficient simulation solutions exist, for example by using a different proposal density for small x).

(c) We can recognize this density as a mixture of normal densities:

$$\pi(x) = \sum_{i=1}^{7} w_i \,\text{Normal}(x; u_i, 1)$$

where $\text{Normal}(x; u_i, 1)$ denotes the value in x of the normal density with expectation u_i and variance 1. To simulate from this density, simulate first an index i according to the probabilities $w_1, \ldots, w_7$. This can be done by simulating U uniformly on $[0, 1]$ and finding the smallest i such that $w_1 + \cdots + w_i \geq U$. Then, if ϕ^{-1} is the inverse of the cumulative distribution function for the standard normal distribution, we may output

$$\phi^{-1}(V) + u_i$$

where V is uniformly simulated on $[0, 1]$.

4. (a) A Bayesian Network is a Directed Acyclic Graph (DAG), for each node i in the network a variable x_i, and for each such node a conditional probability density $\pi(x_i \mid x_{j_1}, \ldots, x_{j_k})$, where $j_1, \ldots, j_k$ are the indices of the parents of node i in the DAG. The product of these conditional probability densities represents the joint probability density for the network.

   (b) A Markov Network is an undirected graph, for each node i in the network a variable x_i, and for each set of nodes with indices $j_1, \ldots, j_k$ such that all nodes in the set are connected in the graph a nonnegative function $\phi(x_{j_1}, \ldots, x_{j_k})$. The product of all the factors represents the (unnormalized) probability density of the network.

   (c) If the value x_i of a node i is observed in a Bayesian network, the probability density for the remaining nodes is obtained as the conditional density given x_i. If the value of node i is set by *intervention* in a causal network with the same structure, the probability density for the remaining nodes is obtained by first removing the conditional density $\pi(x_i \mid x_{j_1}, \ldots, x_{j_i})$ in the product of conditional densities representing the Bayesian network, before conditioning on the value x_i in the remaining product.

   (d) Two nodes i and j are connected in the Markov graph if and only if $\tau_{ij} \neq 0$, where τ_{ij} is the value in the i'th row and j'th column of the precision matrix.

5. If an integral is written as $I = \int f(x)g(x)\,dx$ with $g(x)$ being a probability density, then Monte Carlo integration means making the approximation

$$I = E_g\left[f(x)\right] = \int f(x)g(x)\,dx \approx \frac{1}{n}\sum_{i=1}^{n} f(x_i) = \hat{I}$$

where $x_1, \ldots, x_n$ is a sample from the density $g(x)$. As long as $f(x)$ has finite variance σ^2 when x is distributed according to $g(x)$, the Central Limit Theorem tells us that, for large n, and assuming the sample is a random sample, we have approximately

$$\hat{I} \sim \text{Normal}\left(E_g\left[f(x)\right], \sigma^2/n\right)$$

and this can be used to obtain approximate estimates for $\hat{I} - I$.

6. (a) The posterior density for the model can be written as

$$\begin{aligned}
&\pi(\alpha)\pi(\beta)\prod_{i=1}^{k}\left[\left[\prod_{j=1}^{s}\pi(c_{ij}\mid\lambda_i)\right]\pi(\lambda_i\mid\alpha,\beta)\right] \\
\propto\ &\beta^{5-1}\exp(-2\beta)\prod_{i=1}^{k}\left[\left[\prod_{j=1}^{s}\exp(-\lambda_i)\frac{\lambda_i^{c_{ij}}}{c_{ij}!}\right]\frac{\beta^{\alpha}}{\Gamma(\alpha)}\lambda_i^{\alpha-1}\exp(-\beta\lambda_i)\right]
\end{aligned}$$

Taking the logarithm and removing additive terms not depending on α, β, or $\lambda_1, \ldots, \lambda_k$, we get the log posterior

$$\begin{aligned}
&4\log(\beta) - 2\beta + \sum_{i=1}^{k}\left[\left[\sum_{j=1}^{s} -\lambda_i + c_{ij}\log(\lambda_i)\right] + \alpha\log(\beta) - \log(\Gamma(\alpha)) + (\alpha-1)\log(\lambda_i) - \beta\lambda_i\right] \\
=\ &(4+k\alpha)\log(\beta) - 2\beta - k\log(\Gamma(\alpha)) - (s+\beta)\sum_{i=1}^{k}\lambda_i + \sum_{i=1}^{k}\left(\alpha + \sum_{j=1}^{s} c_{ij} - 1\right)\log(\lambda_i)
\end{aligned}$$

(b) Given a function $f(\theta_1, \ldots, \theta_n)$ proportional to a joint density for the parameters $\theta = (\theta_1, \ldots, \theta_n)$, assume you can derive and simulate from each of the conditional distributions $\pi(x_i \mid x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$, for $i = 1, \ldots, n$. Then Gibbs sampling entails first simulating a vector of parameters $\theta^{(0)}$ from some distribution, followed by, for each t, updating $\theta^{(t)}$ to $\theta^{(t+1)}$ by sequentially simulating from the conditional distributions mentioned above, using updated values for the remaining parameters each time. This can be seen as a version of the Metropolis Hastings algorithm, and thus, under general conditions, the distribution of $\theta^{(t)}$ will approach the original joint density when $t \to \infty$.

(c) In the model above, we see from the loglikelihood of question (a) that, for $i = 1, \ldots, k$,

$$\lambda_i \mid \alpha, \beta, c_{i1}, \ldots, c_{is} \sim \text{Gamma}\left(\alpha + \sum_{j=1}^{s} c_{ij}, \beta + s\right)$$

while

$$\beta \mid \alpha\lambda_1, \ldots, \lambda_k \sim \text{Gamma}\left(5 + k\alpha, 2 + \sum_{i=1}^{k}\lambda_i\right)$$

For α we get

$$\pi(\alpha \mid \beta, \lambda_1, \ldots, \lambda_k) \propto \exp\left(\left(k\log(\beta) + \sum_{i=1}^{k}\log(\lambda_i)\right)\alpha - k\log(\Gamma(\alpha))\right)$$

This can be simulated from using for example rejection sampling.

7. (a) The simplest (but not by a long shot the most efficient) alternative would be to use as proposal density g a uniform distribution on the interval $[-15, 35]$. Its density on this

interval would be $1/50 = 0.02$, and so we can see from the figure that choosing for example $M = 7$, we get $Mg(x) \geq f(x)$, where $f(x)$ is the target density. For each iteration, the algorithm would simulate $x \sim \text{Uniform}[-15, 35]$ and $U \sim \text{Uniform}[0, 0.14]$, and would then reject x unless $U \leq f(x)$.

(b) The simplest possiblity would again be to use a uniform distribution on $[-15, 35]$ as proposal function. The main difference with (a) would be that the chain would contain a number of repeated values; it would "get stuck" for a while at values corresponding to the peaks in the figure. However, one may also use a more tailored proposal function, for example a mixture of three normals, with expectations -6, 6, and 11, respectively, and with respective standard deviations 1, 1, and 4, for example.

(c) With a random walk Metropolis Hastings, you would like the Markov chain to be able to jump from one area of high density to another, i.e., occationally it should jump a length around 8. Thus a normal distribution with standard deviation 4 could be suitable.

(d) If you rescale the proposal function to a much smaller variance, there is a danger that the Markov chain would get stuck in one of the areas of high density for a very long time, as a chain passing the areas of low density would be unlikely. If ou rescale to a much larger variance, one would get the problem that the proposed values would very rarely be accepted, and the chain would be stuck at a single value for that reason.