MA40189 - Solution Sheet Eight

Simon Shaw, s.shaw@bath.ac.uk
https://people.bath.ac.uk/masss/ma40189.html

2021/22 Semester II

1. **Let $X_1, \ldots, X_n$ be exchangeable so that the X_i are conditionally independent given a parameter θ. Suppose that $X_i \,|\, \theta \sim$ Inv-gamma(α, θ), where α is known, and we judge that $\theta \sim$ Gamma(α_0, β_0), where α_0 and β_0 are known.**

   (a) **Show that $\theta \,|\, x \sim$ Gamma(α_n, β_n) where $\alpha_n = \alpha_0 + n\alpha$, $\beta_n = \beta_0 + \sum_{i=1}^n \frac{1}{x_i}$, and $x = (x_1, \ldots, x_n)$.**

   As $X_i \,|\, \theta \sim$ Inv-gamma(α, θ) then

$$\begin{aligned} f(x \,|\, \theta) &= \prod_{i=1}^n \frac{\theta^\alpha}{\Gamma(\alpha)} x_i^{-(\alpha+1)} \exp\left(-\frac{\theta}{x_i}\right) \\ &\propto \theta^{n\alpha} \exp\left(-\theta \sum_{i=1}^n \frac{1}{x_i}\right). \end{aligned}$$

   Similarly, as $\theta \sim$ Gamma(α_0, β_0),

$$f(\theta) \propto \theta^{\alpha_0 - 1} \exp\left(-\theta\beta_0\right).$$

   Hence,

$$\begin{aligned} f(\theta \,|\, x) &\propto f(x \,|\, \theta) f(\theta) \\ &\propto \theta^{(\alpha_0 + n\alpha) - 1} \exp\left\{-\theta\left(\beta_0 + \sum_{i=1}^n \frac{1}{x_i}\right)\right\} \end{aligned}$$

   which we recognise as a kernal of a Gamma$(\alpha_0 + n\alpha, \beta_0 + \sum_{i=1}^n \frac{1}{x_i})$. Thus, $\theta \,|\, x \sim$ Gamma(α_n, β_n) as required.

   (b) **We wish to use the Metropolis-Hastings algorithm to sample from the posterior distribution $\theta \,|\, x$ using a normal distribution with mean θ and chosen variance σ^2 as the symmetric proposal distribution.**

      i. **Suppose that, at time t, the proposed value $\theta^* \leq 0$. Briefly explain why the corresponding acceptance probability is zero for such a θ^* and thus that the sequence of values generated by the algorithm are never less than zero.**

      As the proposal distribution is symmetric then the Metropolis-Hastings algorithm reduces to the Metropolis algorithm with acceptance probability

$$\alpha(\theta^{(t-1)}, \theta^*) = \min\left(1, \frac{f(\theta^* \,|\, x)}{f(\theta^{(t-1)} \,|\, x)}\right)$$

where $\theta \,|\, x \sim \text{Gamma}(\alpha_n, \beta_n)$. Now, assuming $\theta^{(t-1)} > 0$, if $\theta^* \leq 0$ then $f(\theta^* \,|\, x) = 0$ and $\alpha(\theta^{(t-1)}, \theta^*) = 0$ so the move to $\theta^* \leq 0$ is rejected and $\theta^{(t)} = \theta^{(t-1)}$. So, provided $\theta^{(0)} > 0$ then the sequence of values generated by the algorithm are never less than zero.

ii. **Describe how the Metropolis-Hastings algorithm works for this example, giving the acceptance probability in its simplest form.**

The algorithm is performed as follows.

1. Choose a starting point $\theta^{(0)}$ for which $f(\theta^{(0)} \,|\, x) > 0$. This will hold for any $\theta^{(0)} > 0$ as $\theta \,|\, x \sim \text{Gamma}(\alpha_n, \beta_n)$.
2. At time t
   - Sample $\theta^* \sim N(\theta^{(t-1)}, \sigma^2)$.
   - Calculate the acceptance probability

$$\begin{aligned}
\alpha(\theta^{(t-1)}, \theta^*) &= \min\left(1, \frac{f(\theta^* \,|\, x)}{f(\theta^{(t-1)} \,|\, x)}\right) \\
&= \begin{cases} \min\left(1, \frac{(\theta^*)^{\alpha_n - 1} \exp(-\beta_n \theta^*)}{(\theta^{(t-1)})^{\alpha_n - 1} \exp(-\beta_n \theta^{(t-1)})}\right) & \theta^* > 0 \\ 0 & \theta^* \leq 0 \end{cases} \\
&= \begin{cases} \min\left(1, \left(\frac{\theta^*}{\theta^{(t-1)}}\right)^{\alpha_n - 1} \exp\left\{-\beta_n\left(\theta^* - \theta^{(t-1)}\right)\right\}\right) & \theta^* > 0 \\ 0 & \theta^* \leq 0 \end{cases}
\end{aligned}$$

   where $\alpha_n = \alpha_0 + n\alpha$, $\beta_n = \beta_0 + \sum_{i=1}^n \frac{1}{x_i}$.
   - Generate $U \sim U(0, 1)$.
   - If $U \leq \alpha(\theta^{(t-1)}, \theta^*)$ accept the move, $\theta^{(t)} = \theta^*$. Otherwise reject the move, $\theta^{(t)} = \theta^{(t-1)}$.
3. Repeat step 2.

The algorithm will produce a Markov Chain with stationary distribution $f(\theta \,|\, x)$. After a sufficiently long time, b say, to allow for convergence, the values $\{\theta^{(t)}\}$ for $t > b$ may be considered as a sample from $f(\theta \,|\, x)$, that is a sample from Gamma(α_n, β_n), where the samples $\{\theta^{(t)}\}$ for $t \leq b$ are the "burn-in".

2. **Suppose that $X \,|\, \theta \sim N(\theta, \sigma^2)$ and $Y \,|\, \theta, \delta \sim N(\theta - \delta, \sigma^2)$, where σ^2 is a known constant and X and Y are conditionally independent given θ and δ. It is judged that the improper noninformative joint prior distribution $f(\theta, \delta) \propto 1$ is appropriate.**

   (a) **Show that the joint posterior distribution of θ and δ given x and y is bivariate normal with mean vector μ and variance matrix Σ where**

$$\mu = \begin{pmatrix} E(\theta \,|\, X, Y) \\ E(\delta \,|\, X, Y) \end{pmatrix} = \begin{pmatrix} x \\ x - y \end{pmatrix}; \quad \Sigma = \begin{pmatrix} \sigma^2 & \sigma^2 \\ \sigma^2 & 2\sigma^2 \end{pmatrix}.$$

   As $X \,|\, \theta \sim N(\theta, \sigma^2)$ then

$$f(x \,|\, \theta) \propto \exp\left\{-\frac{1}{2\sigma^2}(\theta^2 - 2x\theta)\right\}.$$

As $Y \mid \theta, \delta \sim N(\theta - \delta, \sigma^2)$ then

$$f(y \mid \theta, \delta) \quad \propto \quad \exp\left\{-\frac{1}{2\sigma^2}[(\theta-\delta)^2 - 2y(\theta-\delta)]\right\}.$$

Now, by conditional independence,

$$\begin{aligned} f(x, y \mid \theta, \delta) &= f(x \mid \theta) f(y \mid \theta, \delta) \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}[\theta^2 - 2x\theta + (\theta-\delta)^2 - 2y(\theta-\delta)]\right\}. \end{aligned}$$

So,

$$\begin{aligned} f(\theta, \delta \mid x, y) &\propto f(x, y \mid \theta, \delta) f(\theta, \delta) \\ &\propto f(x, y \mid \theta, \delta) \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}[\theta^2 - 2x\theta + (\theta-\delta)^2 - 2y(\theta-\delta)]\right\} \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}[2\theta^2 - 2(x+y)\theta - 2\theta\delta + \delta^2 + 2y\delta]\right\}. \end{aligned}$$

The exponential argument is quadratic in θ, δ so $\theta, \delta \mid x, y$ is bivariate normal. We can find the exact normal by completing the square or, in this case, using the information given in the question. We have,

$$\Sigma^{-1} \quad = \quad \frac{1}{\sigma^2}\begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}$$

so that the density of the $N_2(\mu, \Sigma)$ is

$$\begin{aligned} &\exp\left\{-\frac{1}{2}\begin{pmatrix}\theta - x & \delta - (x-y)\end{pmatrix}\frac{1}{\sigma^2}\begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} \theta - x \\ \delta - (x-y) \end{pmatrix}\right\} \\ &\quad \propto \quad \exp\left\{-\frac{1}{2\sigma^2}[2\theta^2 - 2(2x - (x-y))\theta - 2\theta\delta + \delta^2 - 2((x-y) - x)\delta]\right\} \\ &\quad = \quad \exp\left\{-\frac{1}{2\sigma^2}[2\theta^2 - 2(x+y)\theta - 2\theta\delta + \delta^2 + 2y\delta]\right\} \\ &\quad \propto \quad f(\theta, \delta \mid x, y). \end{aligned}$$

(b) **Describe how the Gibbs sampler may be used to sample from the posterior distribution $\theta, \delta \mid x, y$, deriving all required conditional distributions.**

The Gibbs sampler requires the conditional distributions $\theta \mid \delta, x, y$ and $\delta \mid \theta, x, y$. Now, with proportionality with respect to θ,

$$\begin{aligned} f(\theta \mid \delta, x, y) &\propto f(\theta, \delta \mid x, y) \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}[2\theta^2 - 2(x+y+\delta)\theta]\right\} \\ &= \exp\left\{-\frac{1}{2}\left(\frac{2}{\sigma^2}\right)\left[\theta^2 - 2\left(\frac{x+y+\delta}{2}\right)\theta\right]\right\} \end{aligned}$$

which is a kernel of a normal density. Thus, $\theta \,|\, \delta, x, y \sim N(\frac{x+y+\delta}{2}, \frac{\sigma^2}{2})$. Now, with proportionality with respect to δ,

$$\begin{aligned} f(\delta \,|\, \theta, x, y) &\propto f(\theta, \delta \,|\, x, y) \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}[\delta^2 - 2(\theta - y)\delta]\right\} \end{aligned}$$

which is a kernel of a normal density. Thus, $\delta \,|\, \theta, x, y \sim N(\theta - y, \sigma^2)$. The Gibbs sampler algorithm is

1. Choose a starting value $(\theta^{(0)}, \delta^{(0)})$ for which $f(\theta^{(0)}, \delta^{(0)} \,|\, x, y) > 0$.
2. At iteration t generate the new values $(\theta^{(t)}, \delta^{(t)})$ as follows:
   - draw $\theta^{(t)}$ from $N(\frac{x+y+\delta^{(t-1)}}{2}, \frac{\sigma^2}{2})$, the distribution of $\theta^{(t)} \,|\, \delta^{(t-1)}, x, y$.
   - draw $\delta^{(t)}$ from $N(\theta^{(t)} - y, \sigma^2)$, the distribution of $\delta^{(t)} \,|\, \theta^{(t)}, x, y$.
3. Repeat step 2.

The algorithm produces a Markov Chain with stationary distribution $f(\theta, \delta \,|\, x, y)$. After a sufficiently long time to allow for convergence, the values $\{\theta^{(t)}, \delta^{(t)}\}$ for $t > b$ may be considered as a sample from $f(\theta, \delta \,|\, x, y)$ where the samples $\{\theta^{(t)}, \delta^{(t)}\}$ for $t \leq b$ are the "burn-in".

(c) **Suppose that $x = 2$, $y = 1$ and $\sigma^2 = 1$. Sketch the contours of the joint posterior distribution. Starting from the origin, add to your sketch the first four steps of a typical Gibbs sampler path.**

For $x = 2$, $y = 1$ and $\sigma^2 = 1$ we shall use the Gibbs sampler to sample from $N_2(\mu, \Sigma)$ where $\mu = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$ and $\Sigma = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$. We draw $\theta^{(t)}$ from $N(\frac{3+\delta^{(t-1)}}{2}, \frac{1}{2})$ and $\delta^{(t)}$ from $N(\theta^{(t)} - 1, 1)$. Suppose that we start from $(1, 1)$ and sample $\theta^{(1)} = 2.216715$, $\delta^{(1)} = 1.475738$, $\theta^{(2)} = 2.086383$ and $\delta^{(2)} = 1.617760$. The path of the trajectory of these points is shown in Figure 1.

(d) **Suppose, instead, that we consider sampling from the posterior distribution using the Metropolis-Hastings algorithm where the proposal distribution is the bivariate normal with mean vector $\tilde{\mu}^{(t-1)} = (\theta^{(t-1)}, \delta^{(t-1)})^T$ and known variance matrix $\tilde{\Sigma}$. Explain the Metropolis-Hastings algorithm for this case, explicitly stating the acceptance ratio.**

For the proposal we have

$$\begin{aligned} &q(\theta^*, \delta^* \,|\, \theta^{(t-1)}, \delta^{(t-1)}) \;\propto \\ &\qquad \exp\left\{-\frac{1}{2}\begin{pmatrix} \theta^* - \theta^{(t-1)} & \delta^* - \delta^{(t-1)} \end{pmatrix} \tilde{\Sigma}^{-1} \begin{pmatrix} \theta^* - \theta^{(t-1)} \\ \delta^* - \delta^{(t-1)} \end{pmatrix}\right\} \end{aligned}$$

so that $q(\theta^*, \delta^* \,|\, \theta^{(t-1)}, \delta^{(t-1)}) = q(\theta^{(t-1)}, \delta^{(t-1)} \,|\, \theta^*, \delta^*)$. The proposal distribution is symmetric and so the Metropolis-Hastings algorithm reduces to the Metropolis algorithm. The algorithm is performed as follows.

1. Choose a starting point $(\theta^{(0)}, \delta^{(0)})$ for which $f(\theta^{(0)}, \delta^{(0)} \,|\, x, y) > 0$.
2. At time t
   - Sample $(\theta^*, \delta^*) \sim N_2(\tilde{\mu}^{(t-1)}, \tilde{\Sigma})$.

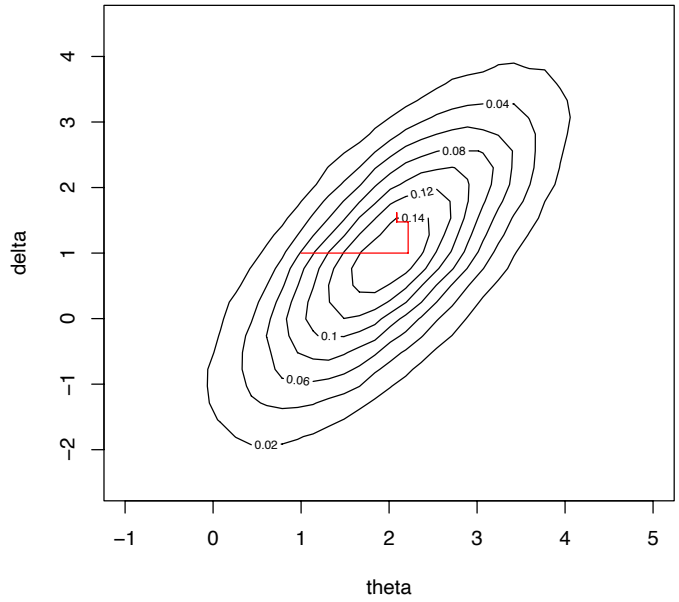

Figure 1: A trajectory path of the Gibbs sampler sampling from $N_2(\mu, \Sigma)$.

- Calculate the acceptance probability

$$\begin{aligned}\alpha((\theta^{(t-1)}, \delta^{(t-1)}), (\theta^*, \delta^*)) &= \min\left(1, \frac{f(\theta^*, \delta^* \,|\, x, y)}{f(\theta^{(t-1)}, \delta^{(t-1)} \,|\, x, y)}\right) \\ &= \min\left(1, \frac{\exp\left\{-\frac{1}{2}(\gamma^* - \mu)^T \Sigma^{-1} (\gamma^* - \mu)\right\}}{\exp\left\{-\frac{1}{2}(\gamma^{(t-1)} - \mu)^T \Sigma^{-1} (\gamma^{(t-1)} - \mu)\right\}}\right)\end{aligned}$$

  where $\gamma^* = \begin{pmatrix} \theta^* \\ \delta^* \end{pmatrix}$ and $\gamma^{(t-1)} = \begin{pmatrix} \theta^{(t-1)} \\ \delta^{(t-1)} \end{pmatrix}$.
- Generate $U \sim U(0, 1)$.
- If $U \leq \alpha((\theta^{(t-1)}, \delta^{(t-1)}), (\theta^*, \delta^*))$ accept the move, $(\theta^{(t)}, \delta^{(t)}) = (\theta^*, \delta^*)$. Otherwise reject the move, $(\theta^{(t)}, \delta^{(t)}) = (\theta^{(t-1)}, \delta^{(t-1)})$.

3. Repeat step 2.

The algorithm produces a Markov Chain with stationary distribution $f(\theta, \delta \,|\, x, y)$. After a sufficiently long time to allow for convergence, the values $\{\theta^{(t)}, \delta^{(t)}\}$ for $t > b$ may be considered as a sample from $f(\theta, \delta \,|\, x, y)$ where the samples $\{\theta^{(t)}, \delta^{(t)}\}$ for $t \leq b$ are the "burn-in".

3. **Let $X_1, \ldots, X_n$ be exchangeable so that the X_i are conditionally independent given a parameter $\theta = (\mu, \lambda)$. Suppose that $X_i \,|\, \theta \sim N(\mu, 1/\lambda)$ so that μ is the mean and λ the precision of the distribution. Suppose that we judge that μ and λ are independent with $\mu \sim N(\mu_0, 1/\tau)$, where μ_0 and τ are known, and $\lambda \sim Gamma(\alpha, \beta)$, where α and β are known.**

(a) **Show that the posterior density $f(\mu, \lambda \,|\, x)$, where $x = (x_1, \ldots, x_n)$, can be expressed as**

$$f(\mu, \lambda \,|\, x) \quad \propto \quad \lambda^{\alpha + \frac{n}{2} - 1} \exp\left\{ -\frac{\lambda}{2} \sum_{i=1}^{n} (x_i - \mu)^2 - \frac{\tau}{2}\mu^2 + \tau\mu_0\mu - \beta\lambda \right\}.$$

As $X_i \,|\, \theta \sim N(\mu, 1/\lambda)$ and $\theta = (\mu, \lambda)$ then

$$\begin{aligned} f(x \,|\, \mu, \lambda) \quad &\propto \quad \prod_{i=1}^{n} \lambda^{\frac{1}{2}} \exp\left\{ -\frac{\lambda}{2}(x_i - \mu)^2 \right\} \\ &= \quad \lambda^{\frac{n}{2}} \exp\left\{ -\frac{\lambda}{2} \sum_{i=1}^{n} (x_i - \mu)^2 \right\}. \end{aligned}$$

$\mu \sim N(\mu_0, 1/\tau)$ so that

$$\begin{aligned} f(\mu) \quad &\propto \quad \exp\left\{ -\frac{\tau}{2}(\mu - \mu_0)^2 \right\} \\ &\propto \quad \exp\left\{ -\frac{\tau}{2}\mu^2 + \tau\mu_0\mu \right\}. \end{aligned}$$

Finally, as $\lambda \sim Gamma(\alpha, \beta)$ then

$$f(\lambda) \quad \propto \quad \lambda^{\alpha - 1} \exp\{-\beta\lambda\}.$$

Hence, as μ and λ are independent,

$$\begin{aligned} f(\mu, \lambda \,|\, x) \quad &\propto \quad f(x \,|\, \mu, \lambda) f(\mu) f(\lambda) \\ &\propto \quad \lambda^{\alpha + \frac{n}{2} - 1} \exp\left\{ -\frac{\lambda}{2} \sum_{i=1}^{n} (x_i - \mu)^2 - \frac{\tau}{2}\mu^2 + \tau\mu_0\mu - \beta\lambda \right\}. \end{aligned}$$

(b) **Hence show that**

$$\lambda \,|\, \mu, x \sim Gamma\left(\alpha + \frac{n}{2}, \beta + \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^2 \right).$$

We have

$$\begin{aligned} f(\lambda \,|\, \mu, x) \quad &= \quad \frac{f(\lambda, \mu \,|\, x)}{f(\mu \,|\, x)} \\ &\propto \quad f(\lambda, \mu \,|\, x) \end{aligned}$$

where the proportionality is with respect to λ. So, from 2.(a) we have

$$\begin{aligned} f(\lambda \,|\, \mu, x) \quad &\propto \quad \lambda^{\alpha + \frac{n}{2} - 1} \exp\left\{ -\frac{\lambda}{2} \sum_{i=1}^{n} (x_i - \mu)^2 - \frac{\tau}{2}\mu^2 + \tau\mu_0\mu - \beta\lambda \right\} \\ &\propto \quad \lambda^{\alpha + \frac{n}{2} - 1} \exp\left\{ -\lambda \left(\beta + \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^2 \right) \right\} \end{aligned}$$

which is a kernel of a $Gamma\left(\alpha + \frac{n}{2}, \beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right)$.

(c) **Given that** $\mu \,|\, \lambda, x \sim N(\frac{\tau\mu_0 + n\lambda\bar{x}}{\tau + n\lambda}, \frac{1}{\tau + n\lambda})$**, where** $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$**, describe how the Gibbs sampler may be used to sample from the posterior distribution** $\mu, \lambda \,|\, x$**. Give a sensible estimate of** $Var(\lambda \,|\, x)$**.**

The Gibbs sampler requires the conditional distributions $\lambda \,|\, \mu, x$ and $\mu \,|\, \lambda, x$ which we have. The algorithm is

1. Choose a starting value $(\mu^{(0)}, \lambda^{(0)})$ for which $f(\mu^{(0)}, \lambda^{(0)} \,|\, x) > 0$.
2. At iteration t generate the new values $(\mu^{(t)}, \lambda^{(t)})$ as follows:
   - draw $\mu^{(t)}$ from $N(\frac{\tau\mu_0 + n\lambda^{(t-1)}\bar{x}}{\tau + n\lambda^{(t-1)}}, \frac{1}{\tau + n\lambda^{(t-1)}})$, the distribution of $\mu^{(t)} \,|\, \lambda^{(t-1)}, x$.
   - draw $\lambda^{(t)}$ from $Gamma\left(\alpha + \frac{n}{2}, \beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu^{(t)})^2\right)$, the distribution of $\lambda^{(t)} \,|\, \mu^{(t)}, x$.
3. Repeat step 2.

The algorithm will produce a Markov Chain with stationary distribution $f(\mu, \lambda \,|\, x)$. After a sufficiently long time to allow for convergence, the values $\{\mu^{(t)}, \lambda^{(t)}\}$ for $t > b$ may be considered as a sample from $f(\mu, \lambda \,|\, x)$ where the samples $\{\mu^{(t)}, \lambda^{(t)}\}$ for $t \leq b$ are the "burn-in".

Ergodic averages converge to the desired expectations under the target distribution so that, for example,

$$\frac{1}{N}\sum_{t=1}^{N} g(\lambda^{(t)}) \quad \rightarrow \quad E(g(\lambda) \,|\, X)$$

as $N \rightarrow \infty$. As this includes the observations prior to convergence, we will do better to estimate only using the observations after the "burn-in". Noting that $Var(\lambda \,|\, X)$ corresponds to $g(\lambda) = (\lambda - E(\lambda \,|\, X))^2$ then any sensible sample variance will suffice. For example, we could estimate $Var(\lambda \,|\, X) = E(\lambda^2 \,|\, X) - E^2(\lambda \,|\, X)$ by

$$\frac{1}{N-b}\sum_{t=b+1}^{N} (\lambda^{(t)} - \bar{\lambda})^2 \quad = \quad \left(\frac{1}{N-b}\sum_{t=b+1}^{N} (\lambda^{(t)})^2\right) - \bar{\lambda}^2$$

where our sample is $\{\mu^{(t)}, \lambda^{(t)} : t = b+1, \ldots, N\}$ and $\bar{\lambda} = \frac{1}{N-b}\sum_{t=b+1}^{N} \lambda^{(t)}$, the sample mean.

4. **Consider a Poisson hierarchical model. At the first stage we have observations** s_j **which are Poisson with mean** $t_j\lambda_j$ **for** $j = 1, \ldots, p$ **where each** t_j **is known. We assume that that the** λ_j **are independent and identically distributed with** $Gamma(\alpha, \beta)$ **prior distributions. The parameter** α **is known but** β **is unknown and is given a** $Gamma(\gamma, \delta)$ **distribution where** γ **and** δ **are known.**

(a) **Find, up to a constant of integration, the joint posterior distribution of the unknown parameters given** $s = (s_1, \ldots, s_p)$**.**

Let $\lambda = (\lambda_1, \ldots, \lambda_p)$ then the unknown parameters are $\theta = (\lambda_1, \ldots, \lambda_p, \beta) = (\lambda, \beta)$. The prior is

$$f(\lambda, \beta) \quad = \quad f(\lambda \,|\, \beta) f(\beta)$$

$$
\begin{aligned}
&= \left\{\prod_{j=1}^{p} f(\lambda_j \mid \beta)\right\} f(\beta) \\
&= \left\{\prod_{j=1}^{p} \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda_j^{\alpha-1} e^{-\beta\lambda_j}\right\} \times \frac{\delta^\gamma}{\Gamma(\gamma)} \beta^{\gamma-1} e^{-\delta\beta} \\
&\propto \left(\prod_{j=1}^{p} \lambda_j^{\alpha-1}\right) \beta^{(p\alpha+\gamma)-1} \exp\left\{-\beta\left(\delta + \sum_{j=1}^{p} \lambda_j\right)\right\}.
\end{aligned}
$$

The likelihood is

$$
\begin{aligned}
f(s \mid \lambda, \beta) &= \prod_{j=1}^{p} \frac{(t_j\lambda_j)^{s_j} e^{-t_j\lambda_j}}{s_j!} \\
&\propto \left(\prod_{j=1}^{p} \lambda_j^{s_j}\right) \exp\left(-\sum_{j=1}^{p} t_j\lambda_j\right).
\end{aligned}
$$

The posterior is thus

$$
\begin{aligned}
f(\lambda, \beta \mid s) &\propto f(s \mid \lambda, \beta) f(\lambda, \beta) \\
&\propto \left(\prod_{j=1}^{p} \lambda_j^{\alpha+s_j-1}\right) \beta^{(p\alpha+\gamma)-1} \exp\left\{-\beta\left(\delta + \sum_{j=1}^{p} \lambda_j\right) - \sum_{j=1}^{p} t_j\lambda_j\right\}.
\end{aligned}
$$

(b) **Describe how the Gibbs sampler may be used to sample from the posterior distribution, deriving all required conditional distributions.**

For each j, let $\lambda_{-j} = \lambda \backslash \lambda_j$. For the Gibbs sampler we need to find the distributions of $\lambda_j \mid \lambda_{-j}, \beta, s$, $j = 1, \ldots, p$, and $\beta \mid \lambda, s$. Now,

$$
\begin{aligned}
f(\lambda_j \mid \lambda_{-j}, \beta, s) &= \frac{f(\lambda, \beta \mid s)}{f(\lambda_{-j}, \beta \mid s)} \\
&\propto f(\lambda, \beta \mid s) \quad \text{(taken with respect to } \lambda_j\text{)} \\
&= \lambda_j^{\alpha+s_j-1} \exp\{-(\beta + t_j)\lambda_j\}
\end{aligned}
$$

which is a kernel of a $Gamma(\alpha + s_j, \beta + t_j)$ distribution. Hence, $\lambda_j \mid \lambda_{-j}, \beta, s \sim Gamma(\alpha + s_j, \beta + t_j)$.

$$
\begin{aligned}
f(\beta \mid \lambda, s) &= \frac{f(\lambda, \beta \mid s)}{f(\lambda \mid s)} \\
&\propto f(\lambda, \beta \mid s) \quad \text{(taken with respect to } \beta\text{)} \\
&= \beta^{(p\alpha+\gamma)-1} \exp\left\{-\left(\delta + \sum_{j=1}^{p} \lambda_j\right)\beta\right\}
\end{aligned}
$$

which is a kernel of a $Gamma(p\alpha + \gamma, \delta + \sum_{j=1}^{p} \lambda_j)$ distribution. So, $\beta \mid \lambda, s \sim Gamma(p\alpha + \gamma, \delta + \sum_{j=1}^{p} \lambda_j)$.

The Gibbs sampler is

1. Choose an arbitrary starting point for which $f(\lambda^{(0)}, \beta^{(0)} \mid s) > 0$. As $\lambda_j > 0$ for $j = 1, \ldots, p$ and $\beta > 0$ then any $p + 1$ collection of positive numbers will suffice.
2. At time t
   - Obtain $\lambda_1^{(t)}$ by sampling from the distribution of $\lambda_1 \mid \lambda_{-1}^{(t-1)}, \beta^{(t-1)}, s$ which is $Gamma(\alpha + s_1, \beta^{(t-1)} + t_1)$.

     $\vdots$
   - Obtain $\lambda_j^{(t)}$ by sampling from the distribution of $\lambda_j \mid \lambda_1^{(t)}, \ldots, \lambda_{j-1}^{(t)}, \lambda_{j+1}^{(t-1)}, \ldots, \lambda_p^{(t-1)}, \beta^{(t-1)}, s$ which is $Gamma(\alpha + s_j, \beta^{(t-1)} + t_j)$.

     $\vdots$
   - Obtain $\lambda_p^{(t)}$ by sampling from the distribution of $\lambda_p \mid \lambda_{-p}^{(t)}, \beta^{(t-1)}, s$ which is $Gamma(\alpha + s_p, \beta^{(t-1)} + t_p)$.
   - Obtain $\beta^{(t)}$ by sampling from the distribution of $\beta \mid \lambda^{(t)}, s$, the $Gamma(p\alpha + \gamma, \delta + \sum_{j=1}^{p} \lambda_j^{(t)})$.
3. Repeat step 2.

The algorithm is run until convergence is reached, When convergence is reached, then resulting value $(\lambda^{(t)}, \beta^{(t)})$ is a realisation from the posterior $\lambda, \beta \mid s$.

(c) **Let $\{\lambda_1^{(t)}, \ldots, \lambda_p^{(t)}, \beta^{(t)}; t = 1, \ldots, N\}$, with N large, be a realisation of the Gibbs sampler described above. Give sensible estimates of $E(\lambda_j \mid s)$, $Var(\beta \mid s)$ and $E(\lambda_j \mid a \leq \beta \leq b, s)$ where $0 < a < b$ are given constants.**

We first remove the burn-in samples. Suppose we determine the burn-in time b, where we assume that the Markov chain has sufficiently converged for $t > b$. (We assume that $b < N$.) Then we discard observations $\{\lambda^{(t)}, \beta^{(t)}; t \leq b\}$ and use the remaining sample $\{\lambda^{(t)}, \beta^{(t)}; t = b + 1, \ldots N\}$ to estimate posterior summaries. Here $\lambda^{(t)} = (\lambda_1^{(t)}, \ldots, \lambda_p^{(t)})$.

The posterior mean of $\lambda_j \mid s$ can be estimated by the sample mean of $\{\lambda_j^{(t)}, t = b + 1, \ldots, N\}$ so that our estimate of $E(\lambda_j \mid s)$ is $\frac{1}{N-b} \sum_{t=b+1}^{N} \lambda_j^{(t)}$.

The posterior variance of $\beta \mid s$ can be estimated by the sample variance of $\{\beta^{(t)}, t = b + 1, \ldots, N\}$ so that our estimate of $Var(\beta \mid s)$ is $\frac{1}{N-b} \sum_{t=b+1}^{N} (\beta^{(t)} - \bar{\beta})^2 = (\frac{1}{N-b} \sum_{t=b+1}^{N} (\beta^{(t)})^2) - \bar{\beta}^2$ where $\bar{\beta} = \frac{1}{N-b} \sum_{t=b+1}^{N} \beta^{(t)}$.

For estimating $E(\lambda_j \mid a \leq \beta \leq b, s)$ we not only discard the burn-in observations but also those elements $\lambda_j^{(t)}$ for which $\beta^{(t)} < a$ or $\beta^{(t)} > b$ for each $t = b + 1, \ldots, N$. So, letting $\mathbb{I}_{[a \leq \beta^{(t)} \leq b]}$ denote the indicator function which is one if $a \leq \beta^{(t)} \leq b$, the number of observations after the burn-in with $a \leq \beta^{(t)} \leq b$ is $M = \sum_{t=b+1}^{N} \mathbb{I}_{[a \leq \beta^{(t)} \leq b]}$ and our estimate of $E(\lambda_j \mid a \leq \beta \leq b, s)$ is $\frac{1}{M} \sum_{t=b+1}^{N} \lambda_j^{(t)} \mathbb{I}_{[a \leq \beta^{(t)} \leq b]}$.

5. **Show that the Gibbs sampler for sampling from a distribution $\pi(\theta)$ where $\theta = (\theta_1, \ldots, \theta_d)$ can be viewed as a special case of the Metropolis-Hastings algorithm where each iteration t consists of d Metropolis-Hastings steps each with an acceptance probability of 1.**

For the Gibbs sampler, we update each θ_p one at a time by obtaining $\theta_p^{(t)}$ from the conditional distribution $\pi(\theta_p \,|\, \theta_1^{(t)}, \ldots, \theta_{p-1}^{(t)}, \theta_{p+1}^{(t-1)}, \ldots, \theta_d^{(t-1)})$. Thus, we move from $\theta_p^{(t-1)}$ to $\theta_p^{(t)}$ leaving the remaining $d-1$ parameters fixed.

We now demonstrate that this update step in the Gibbs sampler is equivalent to a single Metropolis-Hastings step with acceptance probability 1. Let $\theta_{[p]}^{(t-1)} = (\theta_1^{(t)}, \ldots, \theta_{p-1}^{(t)}, \theta_p^{(t-1)}, \theta_{p+1}^{(t-1)}, \ldots, \theta_d^{(t-1)})$ and $\theta_{[p]}^{(t)} = (\theta_1^{(t)}, \ldots, \theta_{p-1}^{(t)}, \theta_p^{(t)}, \theta_{p+1}^{(t-1)}, \ldots, \theta_d^{(t-1)})$. Consider the move from $\theta_{[p]}^{(t-1)}$ to $\theta_{[p]}^{(t)}$. As $\theta_{[p]}^{(t)}$ is drawn from $\pi(\theta_p \,|\, \theta_1^{(t)}, \ldots, \theta_{p-1}^{(t)}, \theta_{p+1}^{(t-1)}, \ldots, \theta_d^{(t-1)})$ then we shall use this for the proposal distribution in the Metropolis-Hastings algorithm so that

$$q(\theta_{[p]}^{(t)} \,|\, \theta_{[p]}^{(t-1)}) \;=\; \pi(\theta_p^{(t)} \,|\, \theta_1^{(t)}, \ldots, \theta_{p-1}^{(t)}, \theta_{p+1}^{(t-1)}, \ldots, \theta_d^{(t-1)}). \tag{1}$$

If, instead, we consider the move from $\theta_{[p]}^{(t)}$ to $\theta_{[p]}^{(t-1)}$ then, using Gibbs sampling, $\theta_{[p]}^{(t-1)}$ is obtained by drawing $\theta_p^{(t-1)}$ from $\pi(\theta_p \,|\, \theta_1^{(t)}, \ldots, \theta_{p-1}^{(t)}, \theta_{p+1}^{(t-1)}, \ldots, \theta_d^{(t-1)})$. Thus, we set

$$q(\theta_{[p]}^{(t-1)} \,|\, \theta_{[p]}^{(t)}) \;=\; \pi(\theta_p^{(t-1)} \,|\, \theta_1^{(t)}, \ldots, \theta_{p-1}^{(t)}, \theta_{p+1}^{(t-1)}, \ldots, \theta_d^{(t-1)}). \tag{2}$$

Now,

$$\begin{aligned}
\pi(\theta_{[p]}^{(t-1)}) \;&=\; \pi(\theta_1^{(t)}, \ldots, \theta_{p-1}^{(t)}, \theta_p^{(t-1)}, \theta_{p+1}^{(t-1)}, \ldots, \theta_d^{(t-1)}) \\
&=\; \pi(\theta_p^{(t-1)} \,|\, \theta_1^{(t)}, \ldots, \theta_{p-1}^{(t)}, \theta_{p+1}^{(t-1)}, \ldots, \theta_d^{(t-1)}) \pi(\theta_1^{(t)}, \ldots, \theta_{p-1}^{(t)}, \theta_{p+1}^{(t-1)}, \ldots, \theta_d^{(t-1)}) \qquad (3) \\
&=\; q(\theta_{[p]}^{(t-1)} \,|\, \theta_{[p]}^{(t)}) \pi(\theta_1^{(t)}, \ldots, \theta_{p-1}^{(t)}, \theta_{p+1}^{(t-1)}, \ldots, \theta_d^{(t-1)}) \qquad (4)
\end{aligned}$$

where (4) follows by substituting (2) into (3). In an identical fashion,

$$\begin{aligned}
\pi(\theta_{[p]}^{(t)}) \;&=\; \pi(\theta_1^{(t)}, \ldots, \theta_{p-1}^{(t)}, \theta_p^{(t)}, \theta_{p+1}^{(t-1)}, \ldots, \theta_d^{(t-1)}) \\
&=\; \pi(\theta_p^{(t)} \,|\, \theta_1^{(t)}, \ldots, \theta_{p-1}^{(t)}, \theta_{p+1}^{(t-1)}, \ldots, \theta_d^{(t-1)}) \pi(\theta_1^{(t)}, \ldots, \theta_{p-1}^{(t)}, \theta_{p+1}^{(t-1)}, \ldots, \theta_d^{(t-1)}) \qquad (5) \\
&=\; q(\theta_{[p]}^{(t)} \,|\, \theta_{[p]}^{(t-1)}) \pi(\theta_1^{(t)}, \ldots, \theta_{p-1}^{(t)}, \theta_{p+1}^{(t-1)}, \ldots, \theta_d^{(t-1)}) \qquad (6)
\end{aligned}$$

where (6) follows by substituting (1) into (5). Dividing (6) by (5) gives

$$\frac{\pi(\theta_{[p]}^{(t)})}{\pi(\theta_{[p]}^{(t-1)})} \;=\; \frac{q(\theta_{[p]}^{(t)} \,|\, \theta_{[p]}^{(t-1)})}{q(\theta_{[p]}^{(t-1)} \,|\, \theta_{[p]}^{(t)})}$$

so that

$$\frac{\pi(\theta_{[p]}^{(t)}) q(\theta_{[p]}^{(t-1)} \,|\, \theta_{[p]}^{(t)})}{\pi(\theta_{[p]}^{(t-1)}) q(\theta_{[p]}^{(t)} \,|\, \theta_{[p]}^{(t-1)})} \;=\; 1. \tag{7}$$

Thus, a Metropolis-Hastings algorithm for sampling from $\pi(\theta)$ for a proposed move from $\theta_{[p]}^{(t-1)}$ to $\theta_{[p]}^{(t)}$ using the proposal distribution given by (1) has an acceptance ratio of

$$\begin{aligned}
\alpha(\theta_{[p]}^{(t-1)}, \theta_{[p]}^{(t)}) \;&=\; \min\left(1, \frac{\pi(\theta_{[p]}^{(t)}) q(\theta_{[p]}^{(t-1)} \,|\, \theta_{[p]}^{(t)})}{\pi(\theta_{[p]}^{(t-1)}) q(\theta_{[p]}^{(t)} \,|\, \theta_{[p]}^{(t-1)})}\right) \qquad (8) \\
&=\; 1 \qquad (9)
\end{aligned}$$

where (9) follows from (8) using (7).