Markov Chain Monte Carlo

Chiranjit Mukhopadhyay
Indian Institute of Science

1 Introduction

As is well known, given a data set $\boldsymbol{Y} = \boldsymbol{y}$, on a random variable $Y \sim f(y|\boldsymbol{\theta})$, all information about the unknown parameter $\boldsymbol{\theta}$ is contained in its posterior distribution

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) \propto L(\boldsymbol{\theta}|\boldsymbol{y})\pi(\boldsymbol{\theta}) \tag{1}$$

where the likelihood function $L(\boldsymbol{\theta}|\boldsymbol{y}) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta})$ is the prior density of $\boldsymbol{\theta}$. Bayesian inference about the model and model parameters proceeds with this posterior distribution, which among other things involve calculation of posterior moments and quantiles, posterior probabilities of $\boldsymbol{A} \subseteq \boldsymbol{\Theta}$, marginal densities of components of $\boldsymbol{\theta}$, predictive densities of Y etc.. All these involve an appropriate integral of the posterior. In most applications it is futile to expect to analytically obtain these integrals. Thus one has to resort to numerical methods. However if $\boldsymbol{\theta}$ is $p \times 1$, for p even as small as 4, brute force numerical integration methods are usually prohibitively time consuming even in today's gazillion instructions per second computing capability due to the so-called "curse of dimension". This led the researchers in search of alternative methods, which is briefly described in the next paragraph.

$\pi(\boldsymbol{\theta}|\boldsymbol{y})$ is a probability density and we are interested in its various features, some examples of which are mentioned in the preceding paragraph. Now even if the study of these exact features may involve integrals of $\pi(\boldsymbol{\theta}|\boldsymbol{y})$, most of these features can be at least approximately studied if one has a large enough sample $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_N$ from this joint posterior density of $\boldsymbol{\theta}$. For instance if one is interested in the component-wise posterior mean of $\boldsymbol{\theta}$, theoretically given by $\int_{\boldsymbol{\Theta}} \boldsymbol{\theta}\pi(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta}$, one can easily approximate this quantity of interest by $\frac{1}{N}\sum_{i=i}^{N} \boldsymbol{\theta}_i$ by virtue of the law of large numbers. If one is interested in the median of θ_j for some $j \in \{1, 2, \ldots, p\}$, it is easily approximated by the sample median calculated from $\theta_{j1}, \theta_{j2} \ldots, \theta_{jN}$. If we are to calculate $\int_{\boldsymbol{A}} \pi(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta}$ for some $\boldsymbol{A} \subseteq \boldsymbol{\Theta}$, for large N, this integral is well approximated by $\frac{1}{N}\sum_{i=i}^{N} I_{\boldsymbol{A}}(\boldsymbol{\theta}_i)$, where $I_A(x)$ is the indicator function of the set A taking values 1 if $x \in A$ and 0 if $x \notin A$. In general, if one is interested in $\int_{\boldsymbol{\Theta}} g(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta}$ for some known function $g(\cdot)$, by the law of large numbers, it is well approximated by $\frac{1}{N}\sum_{i=i}^{N} g(\boldsymbol{\theta}_i)$. Since most of the features of $\phi(\boldsymbol{\theta}|\boldsymbol{y})$ we are interested in Bayesian analysis can be expressed as $\int_{\boldsymbol{\Theta}} g(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{y})\, d\boldsymbol{\theta}$ for some function $g(\cdot)$, this solves the problem of numerically studying them, provided we have a large enough sample $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_N$ from $\pi(\boldsymbol{\theta}|\boldsymbol{y})$.

This leads to devising methods of drawing sample from a possibly high dimensional joint density $\pi(\boldsymbol{\theta}|\boldsymbol{y})$. In the next section we first review some basic techniques of drawing samples or simulating observations from univariate densities, and then later move on to more advanced techniques for multi-dimensional densities involving setting up a Markov Chain (MC) with $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ as its stationary or invariant distribution, and then drawing samples from s this MC using its transition density, which has popularly come to be known as the method of Markov Chain Monte Carlo or MCMC for short.

2 Univariate Methods

In this section we discuss some standard traditional methods of drawing samples from a univariate density $\pi(\theta)$. Though in this section we have suppressed the dependence of $\pi(\cdot)$ on $\boldsymbol{y}$ for the sake of brevity, indeed what we have is a posterior density $\pi(\theta)$ from which we wish to draw samples.

At the outset we assume that we have a method of drawing sample from a $U[0,1]$ distribution, which is the Uniform distribution in $[0,1]$. For instance an algorithm called **linear congruential generator** gives a sequence of pseudo-random integers $\{r_n\}$ between 0 and $M-1$ according to the formula $r_{n+1} = (Ar_n + B)\text{mod}M$ for certain choices of (A, B, M). Using this algorithm one can generate a series of $U[0,1]$ variates. We need not get into the details of this because all programming languages typically provide a library function to generate such $U[0,1]$ variates (like the function `random()` in `C`'s `stdlib.h` for instance). Thus we start discussing the methods of generating observations from a univariate density $\pi(\theta)$, assuming that we can draw samples from a $U[0,1]$ distribution.

2.1 Discrete $\pi(\theta)$

Suppose $\pi(\theta)$ is discrete with support $\{\theta_1, \theta_2, \ldots, \theta_k\}$ and p.m.f. $\{\pi_1, \pi_2, \ldots, \pi_k\}$, with the understanding that $P(\theta = \theta_j) = \pi_j$ for $j = 1, 2, \ldots, k$. In order to generate an observation from this distribution, generate a $U \sim U[0,1]$ and then take the generated θ as θ_j if $\sum_{i=1}^{j-1} \pi_i \leq U < \sum_{i=1}^{j} \pi_i$, where a vacuous sum (occurring for $j = 1$) is defined to be 0. Since $\sum_{i=1}^{k} \pi_i = 1$, for every $U \in [0,1)$ one is guaranteed to find a unique j satisfying the above condition. It is easy to check that the θ generated by the above method has support $\{\theta_1, \theta_2, \ldots, \theta_k\}$ with $P(\theta = \theta_j) = \pi_j$ for $j = 1, 2, \ldots, k$.

2.2 Inversion Method

Suppose θ has p.d.f. $\pi(\theta)$ with c.d.f. $\Pi(\theta) = \int_{-\infty}^{\theta} \pi(\phi)\ d\phi$ such that $\Pi^{-1}(\cdot)$, the inverse function of $\Pi(\theta)$, can at least be numerically obtained. Then in order to generate a θ from this distribution, first generate a $U \sim U[0,1]$ and then set $\theta = \Pi^{-1}(U)$. Since $\Pi(\cdot)$ is a c.d.f. $0 \leq \Pi(\cdot) \leq 1\ \forall\theta$, and thus the domain of $\Pi^{-1}(\cdot)$ is $[0,1]$ so that $\Pi^{-1}(U)$ is a well-defined quantity. A θ generated using this method has the c.d.f. $P(\theta \leq \phi) = P(\Pi^{-1}(U) \leq \phi) = P(U \leq \Pi(\phi) = \Pi(\phi)$, since $U \sim U[0,1]$, showing that the generated θ has the c.d.f. $\Pi(\theta)$ and is thus a sample from the p.d.f. $\pi(\theta)$.

Example 1: Suppose $\theta \sim$exp(λ) so that $\pi(\theta) = \lambda e^{-\lambda\theta}$ and $\Pi(\theta) = 1$ - $e^{-\lambda\theta}$. For this $\Pi(\cdot)$, $\Pi^{-1}(u) = -\left(\log(1-u)\right)/\lambda$. Thus after generating a $U \sim U(0,1)$, set $\theta = -\left(\log(1-U)\right)/\lambda$ and you have an observation from the exp(λ) distribution. Note that since $0 < U < 1$, $0 < 1 - U < 1$ so that $\log(1-U) < 0$ and the generated $\theta > 0$. $\bigtriangledown$

2.3 Transformation Method

Sometimes it is possible to catch hold of a transformation ϕ =$g(\theta)$ such that it is easy to draw sample from the distribution of ϕ. Then after generating a ϕ from this "easy to draw sample from" distribution of ϕ, one applies the inverse transformation $g^{-1}(\phi)$ to get a sample on θ. A couple of examples should drive home the point.

Example 2: Suppose θ ~Weibull(λ, β) so that $\pi(\theta) = \lambda\beta\theta^{\beta-1}e^{-\lambda\theta^{\beta}}$. Now it is easy to show that if θ ~Weibull(λ, β), θ^{β} ~exp(λ). Thus to generate an observation from Weibull(λ, β), first generate an observation ϕ from exp(λ) using the inversion method as in Example 1, and then take $\theta = \phi^{1/\beta}$. Note that in terms of the elementary $U \sim U(0, 1)$, $\theta = \left[-\left(\log(1-U)\right)/\lambda\right]^{1/\beta}$, which is exactly same as inverting the Weibull c.d.f. $\Pi(\theta) = 1$ - $e^{-\lambda\theta^{\beta}}$. ▽

In example 2, the transformation method is essentially equivalent to the same inversion method mentioned in §2.2. To appreciate the transformation method, let us look at the next example.

Example 3: Suppose $\theta \sim N(\mu, \sigma^2)$. The first stage transformation is $\phi = (\theta - \mu)/\sigma$ such that $\phi \sim N(0, 1)$, the standard Normal distribution. Thus if we have a method of generating a ϕ from the standard Normal distribution then we can take $\theta = \mu + \sigma\phi$. But now how can one draw a sample from a $N(0, 1)$ distribution? The c.d.f. of the $N(0, 1)$ distribution is given by $\Phi(z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-\frac{1}{2}x^2}\,dx$, inverting which is not a very easy task with no analytical solution, and thus the inversion method, though can be used in principle, does not readily render a solution. This problem is circumvented by considering the following transformation, called Box-Muller transformation.

Let ϕ_1 and ϕ_2 be i.i.d. $N(0, 1)$. Then their joint density is given by $\frac{1}{2\pi}e^{-\frac{1}{2}(\phi_1^2+\phi_2^2)}$, $-\infty < \phi_1, \phi_2 < \infty$. Now consider a polar transformation of the co-ordinates given by $\phi_1 = r\cos(\psi)$ and $\phi_2 = r\sin(\psi)$. For this transformed variables the range of r is $[0, \infty)$ and that of ψ is $[0, 2\pi]$, and their joint density is given by $\frac{1}{2\pi}re^{-\frac{1}{2}r^2}$. The term r comes from the Jacobian of the transformation given by $\begin{vmatrix} \frac{\partial\phi_1}{\partial r} & \frac{\partial\phi_1}{\partial\psi} \\ \frac{\partial\phi_2}{\partial r} & \frac{\partial\phi_2}{\partial\psi} \end{vmatrix} = \begin{vmatrix} \cos(\psi) & -r\sin(\psi) \\ \sin(\psi) & r\cos(\psi) \end{vmatrix} = r$. Thus from the joint density of (r, ψ), $\frac{1}{2\pi}re^{-\frac{1}{2}r^2}$ for $0 \leq r < \infty$ and $0 \leq \psi < 2\pi$ it is clear that r and ψ are independent with the density of r given by $re^{-\frac{1}{2}r^2}I_{[0,\infty)}(r)$ and that of ψ given by $\frac{1}{2\pi}I_{[0,2\pi]}(\psi)$. Now consider a further transformation of r given by $s = \frac{1}{2}r^2$. Then by a simple change of variable the density of s is given by $e^{-s}I_{[0,\infty)}(s)$, so that s~exp(1). Now for s~exp(1), $e^{-s} \sim U[0, 1]$ and for $\psi \sim U[0, 2\pi]$ $\frac{1}{2\pi}\psi \sim U[0, 1]$. Thus finally we get that if U_1 and U_2 are i.i.d. $U[0, 1]$, write $s = -\log(U_1)$ so that $r = \sqrt{2s} = \sqrt{-2\log(U_1)}$ and $\psi = 2\pi U_2$, such that $\phi_1 = r\cos(\psi) = \sqrt{-2\log(U_1)}\cos(2\pi U_2)$ and $\phi_2 = r\sin(\psi) = \sqrt{-2\log(U_1)}\sin(2\pi U_2)$ are i.i.d $N(0, 1)$. Thus after drawing two independent $U[0, 1]$ variates if one subjects them to the above transformation, one obtains two independent $N(0, 1)$ variates. This yields a method of drawing observations from an arbitrary $N(\mu, \sigma^2)$ distribution. ▽

2.4 Rejection Method

So far we have assumed that the exact form of the p.d.f. $\pi(\theta)$ or the c.d.f. $\Pi(\theta)$ is completely known. But as in (1) many times it is possible that only the form of the posterior is known without any knowledge about the normalizing constant. Thus let the (proper) posterior density $\pi(\theta) = c_f f(\theta)$ where c_f is unknown and $f(\theta)$ is known with $\int_{-\infty}^{\infty} f(\theta)\ d\theta = 1/c_f$ so that $\int_{-\infty}^{\infty} \pi(\theta\ d\theta) = 1$. The problem is to generate an observation from $\pi(\theta)$. Suppose $g(\phi)$ be a density function (*i.e.* $g(\phi) \geq 0$ and $\int_{-\infty}^{\infty} g(\phi)\ d\phi = 1$) such that for some known constant c, $f(\theta) \leq cg(\theta)$ $-\infty < \theta < \infty$, and one can easily draw observations from $g(\phi)$ using one of the methods discussed above. Now in order to generate an observation from $\pi(\theta)$, generate an $U \sim U(0,1)$ and a ϕ from the density $g(\phi)$ independently of each other. Then accept the generated ϕ as an observation θ from $\pi(\theta)$ if $U \leq f(\phi)/\left\{cg(\phi)\right\}$, otherwise again generate a different pair of (U, ϕ) independently of each other till the generated ϕ value is accepted according to the above criterion. Let the final accepted value be denoted by θ. The θ thus generated has the density $\pi(\theta)$ because of the following.

$$\begin{aligned}
& P\left(\phi \leq y \middle| U \leq \frac{f(\phi)}{cg(\phi)}\right) \\
= \quad & \frac{P\left(\phi \leq y, U \leq \frac{f(\phi)}{cg(\phi)}\right)}{P\left(U \leq \frac{f(\phi)}{cg(\phi)}\right)} \\
= \quad & \frac{\int_{-\infty}^{y} P\left(U \leq \frac{f(\phi)}{cg(\phi)} \middle| \phi = z\right) g(z)\ dz}{\int_{-\infty}^{\infty} P\left(U \leq \frac{f(\phi)}{cg(\phi)} \middle| \phi = z\right) g(z)\ dz} \\
= \quad & \frac{\int_{-\infty}^{y} \frac{f(z)}{cg(z)} g(z)\ dz}{\int_{-\infty}^{\infty} \frac{f(z)}{cg(z)} g(z)\ dz} \\
= \quad & \frac{\int_{-\infty}^{y} f(z)\ dz}{\int_{-\infty}^{\infty} f(z)\ dz} \\
= \quad & c_f \int_{-\infty}^{y} f(z)\ dz \\
= \quad & \int_{-\infty}^{y} \pi(z) dz.
\end{aligned}$$

Computationally this method is a major breakthrough which allows one to generate observations from a $\pi(\theta)$ without requiring any knowledge about the normalizing constant. However the price one pays for this is one have to keep on generating (U, ϕ) from $(U[0,1], g(\phi))$ till ϕ satisfies the condition $U \leq f(\phi)/\left\{cg(\phi)\right\}$. Thus the efficiency of this method depends on the **envelope function** $g(\phi)$ and the constant c. The best one can do for the constant c is choose it as $c = \sup_{-\infty<\theta<\infty} \frac{f(\theta)}{g(\theta)}$, but still with this choice one might end up with an undesirable rate of rejection.

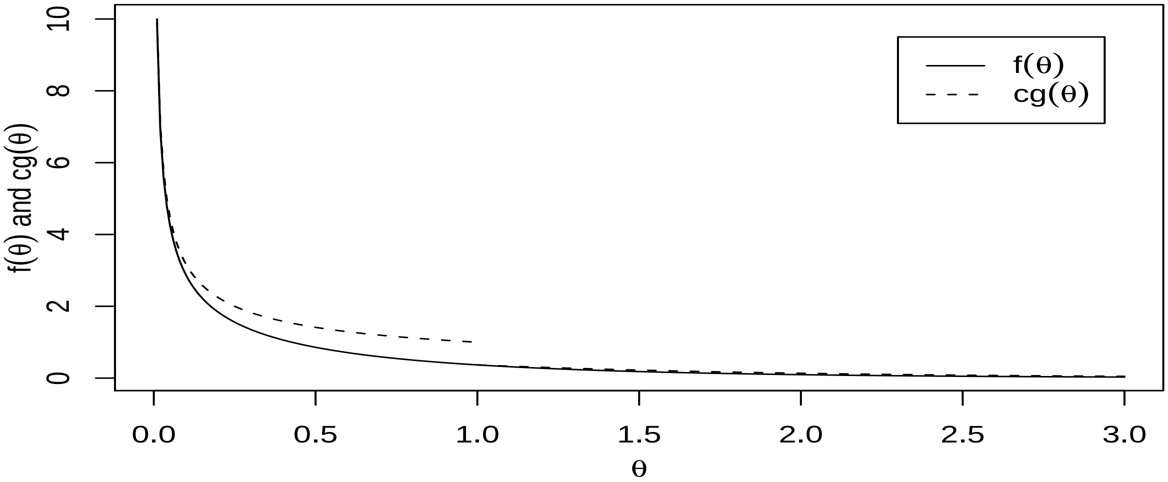
The Rejection method is illustrated in the following important example, which constantly crops up in posterior simulation.

Example 4: Suppose $\theta \sim$Gamma(α, λ) having density $\pi(\theta) \propto \theta^{\alpha-1} e^{-\lambda\theta}$ and we are to generate an observation from this gamma density. First note that if $\psi \sim$Gamma$(\alpha, 1)$ then

$\theta = \psi/\lambda \sim \text{Gamma}(\alpha, \lambda)$. Thus if we can generate an observation ψ from Gamma$(\alpha, 1)$ we immediately have an observation from Gamma(α, λ) as $\theta = \psi/\lambda$. Next notice that if α is a positive integer, then $\psi = \sum_{i=1}^{\alpha} \psi_i$ where the α ψ_i's are i.i.d. exp(1). Thus when α is a positive integer, generate α many i.i.d. exp(1) say $\psi_1, \psi_2, \ldots, \psi_\alpha$ following the method given in Example 1 and then take $\theta = \frac{1}{\lambda}\sum_{i=1}^{\alpha} \psi_i$, which is an observation from the Gamma(α, λ) distribution. Now suppose α is not an integer. Let $a = \alpha - \lfloor \alpha \rfloor$, where $\lfloor \alpha \rfloor$ is the integer part of α or the largest integer $\leq \alpha$. Now $\psi \sim$Gamma$(\alpha, 1)$ has the representation $\psi = \sum_{i=1}^{\lfloor \alpha \rfloor} \psi_i + \psi_a$, where $\psi_1, \psi_2, \ldots, \psi_{\lfloor \alpha \rfloor}$ are i.i.d. exp(1) and ψ_a is independent of the $\lfloor \alpha \rfloor$ ψ_i's and has a Gamma$(a, 1)$ distribution. Thus finally the problem is narrowed down to generating an observation from a Gamma$(a, 1)$ distribution with $0 < a < 1$. This is generated using the rejection method as follows.

Consider the density $g(\phi) = \frac{ae}{a+e}\begin{cases} \phi^{a-1} & \text{if } 0 < \phi < 1 \\ e^{-\phi} & \text{if } \phi \geq 1 \end{cases}$. First note that $g(\phi)$ is a density on $(0, \infty)$ because first of all it is non-negative and $\int_0^\infty g(\phi)\, d\phi = \frac{ae}{a+e}\left[\int_0^1 \phi^{a-1}\, d\phi + \int_1^\infty e^{-\phi}\, d\phi\right]$ $= \frac{ae}{a+e}\left[\frac{1}{a}\, \phi^a\big|_0^1 + (-e^{-\phi})\Big|_1^\infty\right] = \frac{ae}{a+e}\left[\frac{1}{a} + \frac{1}{e}\right] = 1$. The density of Gamma$(a, 1)$ distribution is $\propto \theta^{a-1}e^{-\theta} = f(\theta)$ (say). Now note that for $0 < \theta < 1$, $f(\theta) \leq \theta^{a-1}$ as $e^{-\theta} < 1$; and for $1 \leq \theta < \infty$, $f(\theta) \leq e^{-\theta}$ as $\theta^{a-1} < 1$ since $a < 1$. Thus $\forall \theta \in (0, \infty)$, $f(\theta) \leq \frac{ae}{a+e} g(\theta)$ and also $\frac{f(\theta)}{g(\theta)} = \frac{a+e}{ae}\begin{cases} e^{-\theta} & \text{if } 0 < \theta < 1 \\ \theta^{a-1} & \text{if } \theta \geq 1 \end{cases}$, such that $c = \sup_{-\infty < \theta < \infty} \frac{f(\theta)}{g(\theta)} = \frac{a+e}{ae}$. Thus now we are in a situation where we have the envelope density $g(\phi)$, from which it is easy to draw sample from (as will be seen shortly) and which dominates the target density modulo the normalizing constant $f(\theta)$. This situation is depicted in Figure 1 below for $a = 0.5$.

Figure 1: Envelope Function and Target Density



Thus $f(\theta) \leq cg(\theta)$ $\forall \theta \in (0, \infty)$ where $g(\theta)$ is a proper density. Now the specification will be complete once we explain how to draw an observation from the envelope density $g(\phi)$. This will be done using the inversion method. Thus let $G(\phi) = \int_0^\phi g(\theta)\, d\theta$ denote the c.d.f. of ϕ, which equals $G(\phi) = \begin{cases} \frac{e}{a+e}\phi^a & \text{if } 0 < \phi < 1 \\ \frac{e}{a+e} + \frac{ae}{a+e}(e^{-1} - e^{-\phi}) & \text{if } \phi \geq 1 \end{cases}$. Now in order to generate an observation from this $G(\cdot)$ first generate an $U \sim U(0, 1)$ and then set $\phi =$

$\begin{cases} \left(\frac{a+e}{e}U\right)^{1/a} & \text{if } 0 < U < \frac{e}{a+e} \\ -\log\left(\frac{a+e}{e}\frac{1-U}{a}\right) & \text{if } \frac{e}{a+e} \leq U < 1 \end{cases}$. Thus algorithmically generation of a $\theta \sim$Gamma(α, λ) is as follows.

Step 1. Generate $\psi_1, \psi_2, \ldots, \psi_{\lfloor\alpha\rfloor}$ i.i.d. exp(1) as in Example 1.

Step 2. Generate $U_1 \sim U(0,1)$.

Step 3. Generate $U_2 \sim U(0,1)$ and let $\phi = \begin{cases} \left(\frac{a+e}{e}U_2\right)^{1/a} & \text{if } 0 < U_2 < \frac{e}{a+e} \\ -\log\left(\frac{a+e}{e}\frac{1-U_2}{a}\right) & \text{if } \frac{e}{a+e} \leq U_2 < 1 \end{cases}$, where a $= \alpha$-$\lfloor\alpha\rfloor$.

Step 3. If $\phi < 1$ check if $U_1 \leq e^{-\phi}$ and if $\phi \geq 1$ check if $U_1 \leq \phi^{a-1}$. If the answer is Yes for either case ($\phi < 1$ and $\phi \geq 1$) set $\psi_a = \phi$ and proceed to step 4, otherwise go to Step 2.

Step 4. Let $\theta = \frac{1}{\lambda}\left[\sum_{i=1}^{\lfloor\alpha\rfloor}\psi_i + \psi_a\right]$. $\bigtriangledown$

3 Discrete MC

Let $\{X_n\}_{n=0}^{\infty}$ be a sequence of random variables such that X_n's take value in $\{0, 1, 2, \ldots\}$, called its state space, and

$$P(X_{n+1} = j|X_0 = i_0, X_1 = i_1, X_2 = i_2, \ldots, X_{n-1} = i_{n-1}, X_n = i) = P(X_{n+1} = j|X_n = i) = p_{ij}. \tag{2}$$

If the sequence $\{X_n\}_{n=0}^{\infty}$ satisfies equation (2) then it is called a **Markov Chain** and equation (2) is called the **Markovian** property of the chain. In words, the Markovian property states that given the past $X_0, X_1, \ldots, X_{n-1}$ and the present X_n, the immediate future status X_{n+1} does not depend on the past and it only depends on the present. Also note that this conditional probability $P(X_{n+1} = j|X_n = i)$ dictating the state of immediate future given the present does not depend on n. These p_{ij}'s are called **one step transition probabilities**. Note that each $p_{ij} \geq 0$ and $\sum_{j=0}^{\infty} p_{ij} = 1$ $\forall i = 0, 1, 2, \ldots$. In general we are interested in n-step transition probabilities $p_{ij}^{(n)}$'s denoting $P(X_n = j|X_0 = i)$'s. These are found using the following equations called Chapman-Kolmogorov equations.

$$\begin{aligned} & p_{ij}^{(m+n)} \\ = \; & P(X_{m+n} = j|X_0 = i) \\ = \; & \sum_{k=0}^{\infty} P(X_{m+n} = j, X_m = k|X_0 = i) \\ = \; & \sum_{k=0}^{\infty} P(X_{m+n} = j|X_m = k, X_0 = i)P(X_m = k|X_0 = i) \\ = \; & \sum_{k=0}^{\infty} P(X_{m+n} = j|X_m = k)P(X_m = k|X_0 = i) \\ = \; & \sum_{k=0}^{\infty} P(X_n = j|X_0 = k)p_{ik}^{(m)} \end{aligned}$$

$$= \sum_{k=0}^{\infty} p_{ik}^{(m)} p_{kj}^{(n)}$$

Let $\boldsymbol{P}$ denote the one step transition probability matrix given by $\boldsymbol{P} = \begin{bmatrix} p_{11} & p_{12} & \cdots \\ p_{21} & p_{22} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$. Then by the Chapman-Kolmogorov equation, if one denotes the n-step transition probability matrix by $\boldsymbol{P}^{(n)} = \begin{bmatrix} p_{11}^{(n)} & p_{12}^{(n)} & \cdots \\ p_{21}^{(n)} & p_{22}^{(n)} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$, $\boldsymbol{P}^{(m+n)} = \boldsymbol{P}^{(m)}\boldsymbol{P}^{(n)}$, and therefore $\boldsymbol{P}^{(n)} = \boldsymbol{P}^{(n-1)}\boldsymbol{P}$ (since $\boldsymbol{P}^{(1)} = .\boldsymbol{P}$) $= \boldsymbol{P}^{(n-2)}\boldsymbol{P}^2 = \cdots = \boldsymbol{P}^n$.

In order to get to the limit theorems lending theoretical backbone justifying the MCMC methods we have to first get acquainted with a few definitions.

Two states i and j are said to **communicate** with each other, denoted as $i \leftrightarrow j$ if $\exists m$ and n such that $p_{ij}^{(n)} > 0$ and $p_{Jo}^{(m)} > 0$. It is easy to show that $i \leftrightarrow j$ is an equivalence relation *i.e.* (a) $i \leftrightarrow i$ (b) $i \leftrightarrow j \Rightarrow j \leftrightarrow i$, and (c) $i \leftrightarrow j$ & $j \leftrightarrow k \Rightarrow i \leftrightarrow k$. Thus the states which communicate with each other form disjoint equivalence classes. A MC is said to be **irreducible** if it has only one such equivalence class. That is a chain is irreducible if all of its states communicate with each other.

The **period** of a state i is given by $d(i) = g.c.d\{n : p_{ii}^{(n)} > 0\}$. That is if n is not divisible by $d(i)$ then $p_{ii}^{(n)} = 0$. It is interesting to note that periodicity is class property *i.e.* all states in the same equivalence class induced by the relationship $i \leftrightarrow j$ have the same period. A state is called **aperiodic** if $d(i) = 1$.

State i is called **recurrent** if the probability that the chain comes back to state i at some point of time in the future starting from state i is 1, otherwise it is called **transient**. Like periodicity recurrence can also be shown to be a class property for for equivalence classes induced by $i \leftrightarrow j$.

A probability distribution $\pi_j = P(X = j)$ for $j = 0, 1, 2, \ldots$ is called an **invariant** or **stationary distribution** for a MC if

$$\pi_j = \sum_{i=0}^{\infty} \pi_i p_{ij} \quad \forall j = 0, 1, 2, \ldots \tag{3}$$

It is called so because of the following. Suppose X_0 the starting point of the chain assumes values according to the distribution $\{\pi_i\}_{i=0}^{\infty}$. Then the distribution of X_1 is given by

$$\begin{aligned} & P(X_1 = j) \\ = & \sum_{i=0}^{\infty} P(X_1 = j, X_0 = i) \\ = & \sum_{i=0}^{\infty} P(X_1 = j | X_0 = i) P(X_0 = i) \end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=0}^{\infty} \pi_i p_{ij} \\
&= \pi_j
\end{aligned}
$$

That is in that case the distribution of X_1 is also given by $\{\pi_i\}_{i=0}^{\infty}$ and by induction it follows that the marginal distribution of all the X_n's are also $\{\pi_i\}_{i=0}^{\infty}$. Thus in this case the MC becomes stationary and the distributions of each X_n are invariant.

Now we are in the position to state the main theorem forming the theoretical basis of MCMC computation. Using renewal theory it can be shown that if j is aperiodic $\lim_{n\to\infty} p_{ij}^{(n)}$ exists $\forall i, j$ and the limiting value does not depend on i. If j is transient this limit is always 0. For certain types of recurrent states (called **null recurrent**, which again can be shown to be a class property) also $\lim_{n\to\infty} p_{ij}^{(n)}= 0$. For other types of recurrent states (called **positive recurrent**) this limit is positive and our interest lies in such chains.

Thus consider an irreducible, aperiodic chain. Since it is irreducible, and recurrence/transience and null/positive recurrence are class properties, all its states will fall in one of the following three categories: a) either all states are transient, or b) all states are null recurrent, or c) all states are positive recurrent. We have no interest in cases a) or b) because for these two cases $\lim_{n\to\infty} p_{ij}^{(n)}= 0$ and it can be shown that for such chains there does not exist any invariant distribution. For case c) it can be shown that $\lim_{n\to\infty} p_{ij}^{(n)} = \pi_j$ where π_j is the unique invariant distribution of the chain. If one grants that $\lim_{n\to\infty} p_{ij}^{(n)}$ exists where the limiting value does not depend on i, then it is intuitively very easy to see that these limiting values will yield an invariant distribution. This is because

$$
\begin{aligned}
&\pi_j \\
&= \lim_{n\to\infty} p_{ij}^{(n+1)} \\
&= \lim_{n\to\infty} \sum_{k=0}^{\infty} p_{ik}^{(n)} p_{kj} \\
&= \sum_{k=0}^{\infty} p_{kj} \lim_{n\to\infty} p_{ik}^{(n)} \\
&= \sum_{k=0}^{\infty} \pi_k p_{kj}
\end{aligned}
$$

This result forms the basis of MCMC computation, which may be stated as follows. Start with an aperiodic, irreducible, positive recurrent chain. Start with any initial value $X_0 = i$. Now keep on simulating the next values of the chain using the transition matrix $\boldsymbol{P}$. That is, first generate an X_1 in $\{0, 1, 2, \ldots\}$ using the transition probabilities $\{p_{ij}\}_{j=0}^{\infty}$. Suppose the generated value of X_1 is i_i. Next generate an X_2 in $\{0, 1, 2, \ldots\}$ using the transition probabilities $\{p_{i_1 j}\}_{j=0}^{\infty}$ etc.. In general at the $(n-1)$-th stage if the generated value of X_{n-1}=i_{n-1} then generate an X_n in $\{0, 1, 2, \ldots\}$ using the transition probabilities $\{p_{i_{n-1} j}\}_{j=0}^{\infty}$. Now if one runs this simulation for a long time, say n, and then runs this simulation a large number of times, say N, then according to the above theorem, since $\lim_{n\to\infty} p_{ij}^{(n)} = \pi_j$, the proportion of times X_n in these N simulations equals j, will approximately equal π_j.

This gives one a way to generate a sample from the distribution $\{\pi_j\}_{j=0}^{\infty}$. In the MCMC computation one has this target distribution $\{\pi_j\}_{j=0}^{\infty}$ from which one wants to draw a sample. For this purpose one sets up an aperiodic, irreducible, positive recurrent MC with transition matrix $\boldsymbol{P}$ such that a) $\{\pi_j\}_{j=0}^{\infty}$ is the invariant distribution of $\boldsymbol{P}$, and b) it is easy to sample from $\{0, 1, 2, \ldots\}$ using the probabilities $\{p_{ij}\}_{j=0}^{\infty}$ given in the rows of $\boldsymbol{P}$. Then one simulates this MC as described in the above paragraph in order to obtain a sample from $\{\pi_j\}_{j=0}^{\infty}$.

Practical applications of MCMC involve drawing sample from a p-dimensional joint density $\pi(\boldsymbol{\theta})$ defined on an uncountable state space (typically the p-dimensional Euclidean space $\mathcal{R}^p$) and for this, instead of the transition matrix $\boldsymbol{P}$ one has to appropriately set up transition densities $P(\boldsymbol{\theta}, \boldsymbol{\phi})$ with the interpretation that given $\boldsymbol{X}_n = \boldsymbol{\theta}$, $\boldsymbol{X}_{n+1}$ is distributed according to the conditional density $P(\boldsymbol{\theta}, \boldsymbol{\phi})$=$\lim_{d\boldsymbol{\phi}\to\mathbf{0}} P(\phi_j < X_{n+1,j} < \phi_j + d\phi_j, j = 1, 2, \ldots, p|\boldsymbol{X}_n = \boldsymbol{\theta})/ \prod_{j=1}^{p} d\phi_j$ (where $X_{n,j}$ denotes the j-th component of $\boldsymbol{X}_n$), so that for any $A \subseteq \mathcal{R}^p$, $P(\boldsymbol{X}_{n+1} \in A|\boldsymbol{X}_n = \boldsymbol{\theta})$ denoted by $P(\boldsymbol{\theta}, A)$ can be found as $\int_A P(\boldsymbol{\theta}, \boldsymbol{\phi})d\boldsymbol{\phi}$. MC's with such multi-dimensional state-space and their use in MCMC computation is taken up in the next section.

4 MCMC

Suppose we have a posterior density $\pi(\boldsymbol{\theta})$ with support $\mathcal{R}^p$, from which we wish to sample. As stated in the previous section, for this purpose we shall set up an irreducible, aperiodic, positive recurrent MC with state-space $\mathcal{R}^p$ (and will thus be necessarily vector-valued) and transition density $p(\boldsymbol{\phi}, \boldsymbol{\theta})$ such that this $\pi(\boldsymbol{\theta})$ is going to be the invariant distribution of this chain. Now since we are dealing with a multi-dimensional continuous state-space, the definition of the invariant distribution needs to be suitably modified from that given in (3) for the uni-dimensional discrete case.

$\pi(\boldsymbol{\theta})$ will be called an **invariant distribution** for a MC with state-space $\mathcal{R}^p$ and transition density $p(\boldsymbol{\phi}, \boldsymbol{\theta})$ if

$$\pi(\boldsymbol{\theta}) = \int_{\mathcal{R}^p} p(\boldsymbol{\phi}, \boldsymbol{\theta})\pi(\boldsymbol{\phi})d\boldsymbol{\phi} \quad \forall \boldsymbol{\theta} \in \mathcal{R}^p \tag{4}$$

The interpretation of (4) above is analogous to that of (3). If $\boldsymbol{X}_0$ has density $\pi(\boldsymbol{\phi})$ then since $p(\boldsymbol{\phi}, \boldsymbol{\theta})$ is the conditional density of $\boldsymbol{X}_1$ at $\boldsymbol{\theta}$ given $\boldsymbol{X}_0 = \boldsymbol{\phi}$, $p(\boldsymbol{\phi}, \boldsymbol{\theta})\pi(\boldsymbol{\phi})$ gives the joint density of $(\boldsymbol{X}_0, \boldsymbol{X}_1)$, and thus $\int_{\mathcal{R}^p} p(\boldsymbol{\phi}, \boldsymbol{\theta})\pi(\boldsymbol{\phi})d\boldsymbol{\phi}$ will give the marginal density of $\boldsymbol{X}_1$, which according to (4) is identical to the marginal density of $\boldsymbol{X}_0$. Thus if $\pi(\boldsymbol{\theta})$, the density of $\boldsymbol{X}_0$, satisfies (4), then $\boldsymbol{X}_1$ and inductively $\boldsymbol{X}_2, \boldsymbol{X}_3, \ldots$ all will have marginal density $\pi(\boldsymbol{\theta})$ and hence it is called an invariant distribution of the MC.

Now we shall take up the formalism of n-step transition and the basic limit theorem providing the theoretical back-bone justifying the MCMC computation. Thus given $\boldsymbol{X}_0 = \boldsymbol{\phi}$ we are now interested in the conditional density of $\boldsymbol{X}_n$ at $\boldsymbol{\theta}$ denoted by $p^{(n)}(\boldsymbol{\phi}, \boldsymbol{\theta})$. The expression for $p^{(n)}(\boldsymbol{\phi}, \boldsymbol{\theta})$ is derived following the same logic as $\boldsymbol{P}^{(n)}$ was derived using the Chapman-Kolmogorov equation in the discrete case. The conditional density of $\boldsymbol{X}_n|\boldsymbol{X}_0 = \boldsymbol{\phi}$ at $\boldsymbol{\theta}$ may be obtained by integrating $\boldsymbol{\psi}$, the variable for $\boldsymbol{X}_{n-1}$, out from the joint conditional density of $(\boldsymbol{X}_n, \boldsymbol{X}_{n-1})|\boldsymbol{X}_0 = \boldsymbol{\phi}$ at $(\boldsymbol{\theta}, \boldsymbol{\psi})$. This density is same as the product of the conditional density of $\boldsymbol{X}_n|(\boldsymbol{X}_{n-1} = \boldsymbol{\psi}, \boldsymbol{X}_0 = \boldsymbol{\phi})$ at $\boldsymbol{\theta}$ and the conditional density of $\boldsymbol{X}_{n-1}|\boldsymbol{X}_0 = \boldsymbol{\phi}$ at

$\boldsymbol{\psi}$. The former is nothing but $p(\boldsymbol{\psi}, \boldsymbol{\theta})$ by the Markovian property, while the later is same as $p^{(n-1)}(\boldsymbol{\phi}, \boldsymbol{\psi})$. Thus we get

$$p^{(n)}(\boldsymbol{\phi}, \boldsymbol{\theta}) = \int_{\mathcal{R}^p} p^{(n-1)}(\boldsymbol{\phi}, \boldsymbol{\psi}) p(\boldsymbol{\psi}, \boldsymbol{\theta}) d\boldsymbol{\psi} \tag{5}$$

Intuitively, starting from $\boldsymbol{\phi}$ in order to transit to $\boldsymbol{\theta}$ in n steps, one has to first come somewhere $\boldsymbol{\psi} \in \mathcal{R}^p$ at the $(n-1)$-th step from where you transit to the destination $\boldsymbol{\theta}$ in the next step. The integration allows all possible stepping stone $\boldsymbol{\psi}$ at the $(n-1)$-th step.

While equations (4) and (5) are fine for MC's with continuous state-space $\mathcal{R}^p$, but since $p(\boldsymbol{\phi}, \boldsymbol{\theta})$ is a density, it does not allow a transition from the current value $\boldsymbol{X}_n = \boldsymbol{\phi}$ to itself characterized by $\boldsymbol{X}_{n+1} = \boldsymbol{\phi}$, with a positive probability. But one of the MCMC techniques called the Metropolis-Hastings algorithm has such a feature *i.e.* it allows a transition $\boldsymbol{\phi} \to \boldsymbol{\phi}$ with positive probability. Thus in order to accommodate such movements, equations (4) and (5) are modified as follows.

Let $P(\boldsymbol{\phi}, d\boldsymbol{\theta})$, called the transition kernel, denote the approximate probability of transition from $\boldsymbol{\phi} \to (\boldsymbol{\theta}, \boldsymbol{\theta} + d\boldsymbol{\theta})$, a neighborhood of $\boldsymbol{\theta}$, and likewise let $\pi^*(d\boldsymbol{\theta})$, the invariant distribution, denote the approximate probability of $\boldsymbol{X}_n \in (\boldsymbol{\theta}, \boldsymbol{\theta} + d\boldsymbol{\theta})$, so that in case of densities, $P(\boldsymbol{\phi}, d\boldsymbol{\theta}) \approx p(\boldsymbol{\phi}, \boldsymbol{\theta}) d\boldsymbol{\theta}$ and $\pi^*(d\boldsymbol{\theta}) \approx \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}$. Note that typically $\pi^*(d\boldsymbol{\theta})$ will have the density $\pi(\boldsymbol{\theta})$ but $P(\boldsymbol{\phi}, d\boldsymbol{\theta})$ need not necessarily have an associated transition density. Now with these notations, which would allow a degenerate transition $\boldsymbol{\phi} \to \boldsymbol{\phi}$, $\pi^*(d\boldsymbol{\theta})$ will be called the invariant distribution if

$$\pi^*(d\boldsymbol{\theta}) = \int_{\mathcal{R}^p} P(\boldsymbol{\phi}, d\boldsymbol{\theta}) \pi^*(d\boldsymbol{\phi}) \quad \forall \boldsymbol{\theta} \in \mathcal{R}^p, \tag{6}$$

and the n-step transition kernel $P^{(n)}(\boldsymbol{\phi}, d\boldsymbol{\theta})$ denoting the approximate probability of $\boldsymbol{X}_n \in (\boldsymbol{\theta}, \boldsymbol{\theta} + d\boldsymbol{\theta})$ given that the chain started at $\boldsymbol{X}_0 = \boldsymbol{\phi}$ is given by

$$P^{(n)}(\boldsymbol{\phi}, d\boldsymbol{\theta}) = \int_{\mathcal{R}^p} P^{(n-1)}(\boldsymbol{\phi}, d\boldsymbol{\psi}) P(\boldsymbol{\psi}, d\boldsymbol{\theta}). \tag{7}$$

Note that how (6) boils down to (4), and (7) reduces to (5) in case of the transition densities with $P(\boldsymbol{\phi}, d\boldsymbol{\theta}) \approx p(\boldsymbol{\phi}, \boldsymbol{\theta}) d\boldsymbol{\theta}$. From this point on till the end of §4.1 we shall work with these generalized versions (6) and (7) rather than the restrictive density versions (4) and (5).

In this general set-up let us first examine why MCMC works, which has already been explained in the last but two paragraphs of §3. The basic reason for the validity of MCMC is that $\lim_{n\to\infty} P^{(n)}(\boldsymbol{\phi}, d\boldsymbol{\theta})$ exists $\forall \boldsymbol{\phi}, \boldsymbol{\theta} \in \mathcal{R}^p$, this limit does not depend on the initial state $\boldsymbol{\phi}$ and equals $\pi^*(d\boldsymbol{\theta})$, the unique invariant distribution of an irreducible, aperiodic, positive recurrent chain. First note that since the chain is irreducible, all states would have the same periodicity and recurrence status *i.e.* either all states are transient or null recurrent or positive recurrent. Not much intuition can be given why the limit should exist except that if the chain is not aperiodic, $P^{(n)}(\boldsymbol{\phi}, d\boldsymbol{\theta})$ would be 0 unless n is a multiple of $d > 1$, the period of the chain, and thus being aperiodic is a necessary condition for the existence of the limit. Actually the limit exists $\forall \boldsymbol{\theta}$. For transient states since there is a positive probability of the chain never coming back to it, and for null recurrent states since the expected waiting time of return is ∞, it is intuitively quite obvious that for such states $\lim_{n\to\infty} P^{(n)}(\boldsymbol{\phi}, d\boldsymbol{\theta})$ should equal 0. Now if we grant that the limit would exist for positive recurrent states, let us next verify the validity of the next two points. Since by the Markovian property, probabilistically the chain restarts itself once it comes back to $\boldsymbol{\theta}$, which it is guaranteed to do in finite time since $\boldsymbol{\theta}$ is positive recurrent, it is intuitively clear why the limit does not depend on the initial state $\boldsymbol{\phi}$ and only depends on the probabilistic nature of the final state $\boldsymbol{\theta}$. Thus let us define $\lim_{n\to\infty} P^{(n)}(\boldsymbol{\phi}, d\boldsymbol{\theta})$ as some quantity $\pi^*(d\boldsymbol{\theta})$, which does not depend on $\boldsymbol{\phi}$. Next let us verify why $\pi^*(d\boldsymbol{\theta})$ should be an invariant distribution for the chain.

$$\begin{aligned}
&\pi^*(d\boldsymbol{\theta}) && \\
&= \lim_{n\to\infty} P^{(n)}(\boldsymbol{\phi}, d\boldsymbol{\theta}) && \text{(by definition of } \pi^*(d\boldsymbol{\theta})\text{)} \\
&= \lim_{n\to\infty} \int_{\mathcal{R}^p} P^{(n-1)}(\boldsymbol{\phi}, d\boldsymbol{\psi}) P(\boldsymbol{\psi}, d\boldsymbol{\theta}) && \text{(by (7))} \\
&= \int_{\mathcal{R}^p} \lim_{n\to\infty} P^{(n-1)}(\boldsymbol{\phi}, d\boldsymbol{\psi}) P(\boldsymbol{\psi}, d\boldsymbol{\theta}) && \text{(assuming that we can interchange the} \\
& && \text{limit and the integral)} \\
&= \int_{\mathcal{R}^p} P(\boldsymbol{\psi}, d\boldsymbol{\theta}) \pi^*(d\boldsymbol{\psi}) && \text{(by definition of } \pi^*(d\boldsymbol{\psi})\text{)}
\end{aligned}$$

showing that the limit must satisfy (6) and thus must be the invariant distribution of the chain. Now we shall be through once we can show that the invariant distribution of a chain is unique. Thus let $\pi'(d\boldsymbol{\theta})$ be another invariant distribution. Then

$$\begin{aligned}
&\pi'(d\boldsymbol{\theta}) && \\
&= \int_{\mathcal{R}^p} P(\boldsymbol{\phi}, d\boldsymbol{\theta}) \pi'(d\boldsymbol{\phi}) && \text{(by (6))} \\
&= \int_{\mathcal{R}^p} P(\boldsymbol{\phi}, d\boldsymbol{\theta}) \left\{ \int_{\mathcal{R}^p} P(\boldsymbol{\psi}, d\boldsymbol{\phi}) \pi'(d\boldsymbol{\psi}) \right\} && \text{(applying (6) again for } \pi'(d\boldsymbol{\phi})\text{)} \\
&= \int_{\mathcal{R}^p} \left\{ \int_{\mathcal{R}^p} P(\boldsymbol{\psi}, d\boldsymbol{\phi}) P(\boldsymbol{\phi}, d\boldsymbol{\theta}) \right\} \pi'(d\boldsymbol{\psi}) && \text{(by interchanging the integrals} \\
& && \text{w.r.t } d\boldsymbol{\phi} \text{ and } d\boldsymbol{\psi}\text{)} \\
&= \int_{\mathcal{R}^p} P^{(2)}(\boldsymbol{\psi}, d\boldsymbol{\theta}) \pi'(d\boldsymbol{\psi}) && \text{(by (7))} \\
&\cdots\cdots\cdots\cdots\cdots\cdots && \\
&\cdots\cdots\cdots\cdots\cdots\cdots && \\
&= \int_{\mathcal{R}^p} P^{(n)}(\boldsymbol{\psi}, d\boldsymbol{\theta}) \pi'(d\boldsymbol{\psi}) && \text{(by repeating similar argument as above)}
\end{aligned}$$

Now letting $n \to \infty$, interchanging the limit and the integral and using the fact that $\lim_{n\to\infty} P^{(n)}(\boldsymbol{\psi},\, d\boldsymbol{\theta}) = \pi^*(d\boldsymbol{\theta})$ we get that

$$\pi'(d\boldsymbol{\theta}) = \int_{\mathcal{R}^p} \pi^*(d\boldsymbol{\theta}) \pi'(d\boldsymbol{\phi}) = \pi^*(d\boldsymbol{\theta})$$

showing that the invariant distribution of the chain is unique, which is given by $\lim_{n\to\infty} P^{(n)}(\boldsymbol{\phi},\, d\boldsymbol{\theta}) = \pi^*(d\boldsymbol{\theta})$. This completes the arguments justifying the MCMC computation. Given a target density $\pi(\boldsymbol{\theta})$, if one can set-up an MC with a transition kernel $P(\boldsymbol{\phi}, d\boldsymbol{\theta})$ such that $\pi^*(d\boldsymbol{\theta}) \approx \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}$ is its invariant distribution satisfying (6), then because of the fact that $\lim_{n\to\infty} P^{(n)}(\boldsymbol{\phi}, d\boldsymbol{\theta}) = \pi^*(d\boldsymbol{\theta})$, one can simulate an observation from $\pi(\boldsymbol{\theta})$ by simulating the chain for a long time using the transition kernel $P(\boldsymbol{\phi}, d\boldsymbol{\theta})$ starting with any initial value $\boldsymbol{\phi}$.

Now on the surface though it appears that finding an appropriate transition kernel $P(\boldsymbol{\phi}, d\boldsymbol{\theta})$ for a given $\pi^*(d\boldsymbol{\theta})$ as its invariant distribution, which is also easy to sample from, is like the proverbial search of a needle in the haystack, fortunately there are two algorithms which allow us to do just that. The first one is called the Metropolis-Hastings algorithm and the second one is called Gibbs sampling. We take these up in the next two sub-sections.

4.1 Metropolis-Hastings Algorithm

Consider the transition kernel

$$P(\boldsymbol{\phi}, d\boldsymbol{\theta}) = p(\boldsymbol{\phi}, \boldsymbol{\theta})d\boldsymbol{\theta} + r(\boldsymbol{\phi})\delta_{\boldsymbol{\phi}}(d\boldsymbol{\theta}) \tag{8}$$

where $p(\boldsymbol{\phi}, \boldsymbol{\theta}) \geq 0$, $p(\boldsymbol{\phi}, \boldsymbol{\phi}) = 0$, $\delta_{\boldsymbol{\phi}}(d\boldsymbol{\theta}) = \begin{cases} 1 & \text{if } \boldsymbol{\phi} \in (\boldsymbol{\theta}, \boldsymbol{\theta} + d\boldsymbol{\theta}) \\ 0 & \text{otherwise} \end{cases}$ and $r(\boldsymbol{\phi}) = 1 - \int_{\mathcal{R}^p} p(\boldsymbol{\phi}, \boldsymbol{\theta})d\boldsymbol{\theta}$. Note that $p(\boldsymbol{\phi}, \boldsymbol{\theta})$ is not necessarily a density, if it is, $r(\boldsymbol{\phi})$=0, but in general we want to allow $r(\boldsymbol{\phi}) > 0$ which requires $\int_{\mathcal{R}^p} p(\boldsymbol{\phi}, \boldsymbol{\theta})d\boldsymbol{\theta} < 1$. The transition kernel in (8) states that given $\boldsymbol{\phi}$, in the next step there is a positive probability $r(\boldsymbol{\phi})$ that the chain stays at $\boldsymbol{\phi}$ and the probability of moving to the neighborhood of some other value $\boldsymbol{\theta}$ is $\approx p(\boldsymbol{\phi}, \boldsymbol{\theta})d\boldsymbol{\theta}$. Note that the transition kernel is completely specified by the function $p(\boldsymbol{\phi}, \boldsymbol{\theta})$ and $\int_{\mathcal{R}^p} P(\boldsymbol{\phi}, d\boldsymbol{\theta}) = 1$. Now a sufficient condition for the posterior $\pi(\boldsymbol{\theta})$ to be the invariant density of this transition kernel is given by the condition

$$\pi(\boldsymbol{\theta})p(\boldsymbol{\theta}, \boldsymbol{\phi}) = \pi(\boldsymbol{\phi})p(\boldsymbol{\phi}, \boldsymbol{\theta}) \quad \forall \boldsymbol{\theta}, \boldsymbol{\phi} \in \mathcal{R}^p. \tag{9}$$

Condition (9) is called the reversibility condition. Intuitively, since the l.h.s of (9) gives the unconditional probability of transition from $\boldsymbol{\theta} \to \boldsymbol{\phi}$ and the r.h.s gives the unconditional probability of transition from $\boldsymbol{\phi} \to \boldsymbol{\theta}$, for reversible chains both the (unconditional) probabilities of the transitions from $\boldsymbol{\theta} \to \boldsymbol{\phi}$ and $\boldsymbol{\phi} \to \boldsymbol{\theta}$ are same. Now let us see why condition (9) is sufficient for $\pi(\boldsymbol{\theta})$ to be the invariant density of the transition kernel $P(\boldsymbol{\phi}, d\boldsymbol{\theta})$ given in (8). For $\pi(\boldsymbol{\theta})$ to be invariant for $P(\boldsymbol{\phi}, d\boldsymbol{\theta})$ it has to satisfy the condition given in (6). Thus let us start with the r.h.s. of (6).

$$\begin{aligned}
&\int_{\mathcal{R}^p} P(\boldsymbol{\phi}, d\boldsymbol{\theta})\pi^*(d\boldsymbol{\phi}) \\
&= \left[\int_{\mathcal{R}^p} \pi(\boldsymbol{\phi})p(\boldsymbol{\phi}, \boldsymbol{\theta})d\boldsymbol{\phi}\right] d\boldsymbol{\theta} + \int_{\mathcal{R}^p} \pi(\boldsymbol{\phi})r(\boldsymbol{\phi})\delta_{\boldsymbol{\phi}}(d\boldsymbol{\theta})d\boldsymbol{\phi} \\
&= \left[\int_{\mathcal{R}^p} \pi(\boldsymbol{\theta})p(\boldsymbol{\theta}, \boldsymbol{\phi})d\boldsymbol{\phi}\right] d\boldsymbol{\theta} + \int_{\boldsymbol{\theta}}^{\boldsymbol{\theta}+d\boldsymbol{\theta}} \pi(\boldsymbol{\phi})r(\boldsymbol{\phi})d\boldsymbol{\phi} && \text{(by (9) and definition of } \delta_{\boldsymbol{\phi}}(d\boldsymbol{\theta})) \\
&\approx \left[\int_{\mathcal{R}^p} p(\boldsymbol{\theta}, \boldsymbol{\phi})d\boldsymbol{\phi}\right] \pi(\boldsymbol{\theta})d\boldsymbol{\theta} + \pi(\boldsymbol{\theta})r(\boldsymbol{\theta})d\boldsymbol{\theta} \\
&= (1 - r(\boldsymbol{\theta}))\pi(\boldsymbol{\theta})d\boldsymbol{\theta} + \pi(\boldsymbol{\theta})r(\boldsymbol{\theta})d\boldsymbol{\theta} && \text{(by definition of } r(\boldsymbol{\theta})) \\
&= \pi^*(d\boldsymbol{\theta})
\end{aligned}$$

showing that (9) is sufficient for $\pi(\boldsymbol{\theta})$ to be invariant for $P(\boldsymbol{\phi}, d\boldsymbol{\theta})$.

Now again in general, given $\pi(\boldsymbol{\theta})$ it is not easy to come up with a $p(\boldsymbol{\phi}, \boldsymbol{\theta})$ satisfying (9). But suppose one starts with an arbitrary conditional density $q(\boldsymbol{\phi}, \boldsymbol{\theta})$ with $\int_{\mathcal{R}^p} q(\boldsymbol{\phi}, \boldsymbol{\theta})d\boldsymbol{\theta}$=1. If this $q(\cdot, \cdot)$ exactly satisfies (9), well and good, take $r(\boldsymbol{\phi}) = 0$ and proceed with the transition kernel (8) with $p(\cdot, \cdot)$ replaced by $q(\cdot, \cdot)$. Otherwise suppose for some $(\boldsymbol{\phi}, \boldsymbol{\theta})$

$$\pi(\boldsymbol{\theta})q(\boldsymbol{\theta}, \boldsymbol{\phi}) > \pi(\boldsymbol{\phi})q(\boldsymbol{\phi}, \boldsymbol{\theta}). \tag{10}$$

In this case the probability of transition $\boldsymbol{\theta} \to \boldsymbol{\phi}$ is more than the probability of transition $\boldsymbol{\phi} \to \boldsymbol{\theta}$. Thus to compensate for this imbalance, in such cases we will not always make the transition from $\boldsymbol{\theta} \to \boldsymbol{\phi}$, while will always make the transition $\boldsymbol{\phi} \to \boldsymbol{\theta}$. This notion is crystallized by borrowing idea from the rejection method described in §2.4. In case of the inequality (10), we shall make the transition $\boldsymbol{\theta} \to \boldsymbol{\phi}$ with some probability $\alpha(\boldsymbol{\theta}, \boldsymbol{\phi}) < 1$ and make the transition $\boldsymbol{\phi} \to \boldsymbol{\theta}$ with probability $\alpha(\boldsymbol{\phi}, \boldsymbol{\theta}) = 1$. Thus in general a transition $\boldsymbol{\theta} \to \boldsymbol{\phi}$ is made according to the transition kernel (8) with its $p(\boldsymbol{\theta}, \boldsymbol{\phi})= q(\boldsymbol{\theta}, \boldsymbol{\phi})\alpha(\boldsymbol{\theta}, \boldsymbol{\phi})$, where $\alpha(\cdot, \cdot)$ is determined in such a way that this $p(\boldsymbol{\theta}, \boldsymbol{\phi})$ now satisfies the reversibility condition (9) *i.e.* in case of (10), set $\alpha(\boldsymbol{\phi}, \boldsymbol{\theta}) = 1$ and choose $\alpha(\boldsymbol{\theta}, \boldsymbol{\phi})$ such that

$$\pi(\boldsymbol{\theta})q(\boldsymbol{\theta}, \boldsymbol{\phi})\alpha(\boldsymbol{\theta}, \boldsymbol{\phi}) = \pi(\boldsymbol{\phi})q(\boldsymbol{\phi}, \boldsymbol{\theta})\alpha(\boldsymbol{\phi}, \boldsymbol{\theta}) = \pi(\boldsymbol{\phi})q(\boldsymbol{\phi}, \boldsymbol{\theta}) \Rightarrow \alpha(\boldsymbol{\theta}, \boldsymbol{\phi}) = \frac{\pi(\boldsymbol{\phi})q(\boldsymbol{\phi}, \boldsymbol{\theta})}{\pi(\boldsymbol{\theta})q(\boldsymbol{\theta}, \boldsymbol{\phi})}.$$

Thus in general if we set $\alpha(\boldsymbol{\theta}, \boldsymbol{\phi})=\text{Minimum}\left\{\frac{\pi(\boldsymbol{\phi})q(\boldsymbol{\phi}, \boldsymbol{\theta})}{\pi(\boldsymbol{\theta})q(\boldsymbol{\theta}, \boldsymbol{\phi})}, 1\right\}$, $p(\boldsymbol{\theta}, \boldsymbol{\phi}) = q(\boldsymbol{\theta}, \boldsymbol{\phi})\alpha(\boldsymbol{\theta}, \boldsymbol{\phi})$ satisfies (9), and thus $\pi(\boldsymbol{\theta})$ becomes the invariant distribution of (8) with $q(\boldsymbol{\phi}, \boldsymbol{\theta})\alpha(\boldsymbol{\phi}, \boldsymbol{\theta})$ as its $p(\boldsymbol{\phi}, \boldsymbol{\theta})$.

Above arguments provide the essential logic motivating the development of the transition kernel of the Metropolis-Hastings algorithm, such that the target distribution $\pi(\boldsymbol{\theta})$ becomes the invariant distribution of the developed transition kernel. But now let us turn the table and look at the definition of this developed transition kernel and then check why it works *i.e.* why $\pi(\boldsymbol{\theta})$ is its invariant distribution. This is sort of redundant in view of the above arguments leading to the development of the following transition kernel (11), but nevertheless this recapitulation would hopefully be helpful for readers seeing this algorithm for the first time.

The Metropolis-Hastings transition kernel is given by

$$P(\boldsymbol{\phi}, d\boldsymbol{\theta}) = q(\boldsymbol{\phi}, \boldsymbol{\theta})\alpha(\boldsymbol{\phi}, \boldsymbol{\theta})d\boldsymbol{\theta} + r(\boldsymbol{\phi})\delta_{\boldsymbol{\phi}}(d\boldsymbol{\theta}) \tag{11}$$

where $q(\boldsymbol{\phi}, \boldsymbol{\theta}) \geq 0$, $q(\boldsymbol{\phi}, \boldsymbol{\phi}) = 0$, $\int_{\mathcal{R}^p} q(\boldsymbol{\phi}, \boldsymbol{\theta})d\boldsymbol{\theta} = 1$, $\alpha(\boldsymbol{\phi}, \boldsymbol{\theta}) = \text{Minimum}\left\{\frac{\pi(\boldsymbol{\theta})q(\boldsymbol{\theta}, \boldsymbol{\phi})}{\pi(\boldsymbol{\phi})q(\boldsymbol{\phi}, \boldsymbol{\theta})}, 1\right\}$, $r(\boldsymbol{\phi}) = 1 - \int_{\mathcal{R}^p} q(\boldsymbol{\phi}, \boldsymbol{\theta})\alpha(\boldsymbol{\phi}, \boldsymbol{\theta})d\boldsymbol{\theta}$ and $\delta_{\boldsymbol{\phi}}(d\boldsymbol{\theta}) = \begin{cases} 1 & \text{if } \boldsymbol{\phi} \in (\boldsymbol{\theta}, \boldsymbol{\theta} + d\boldsymbol{\theta}) \\ 0 & \text{otherwise} \end{cases}$. This kernel says that starting at some $\boldsymbol{\phi}$ the probability of transiting to the neighborhood of some $\boldsymbol{\theta} \neq \boldsymbol{\phi}$ is $\approx q(\boldsymbol{\phi}, \boldsymbol{\theta})\alpha(\boldsymbol{\phi}, \boldsymbol{\theta})d\boldsymbol{\theta}$, where $d\boldsymbol{\theta}$ is the hyper-volume of the neighborhood. Since $q(\cdot, \boldsymbol{\theta})$ is a conditional density and $0 \leq \alpha(\boldsymbol{\phi}, \boldsymbol{\theta}) \leq 1$, this transition from $\boldsymbol{\phi} \to \boldsymbol{\theta} \neq \boldsymbol{\phi}$ may be interpreted as first generating a $\boldsymbol{\theta}$ using the conditional density $q(\boldsymbol{\phi}, \boldsymbol{\theta})$ and then making the transition $\boldsymbol{\phi} \to \boldsymbol{\theta}$ with probability $\alpha(\boldsymbol{\phi}, \boldsymbol{\theta})$. Otherwise the chain transits from $\boldsymbol{\phi} \to \boldsymbol{\phi}$, the same value $\boldsymbol{\phi}$, with probability $r(\boldsymbol{\phi})$.

First let us check that the $P(\boldsymbol{\phi}, d\boldsymbol{\theta})$ given in (11) is indeed a legitimate transition kernel by checking that

$$\int_{\mathcal{R}^p} P(\boldsymbol{\phi}, d\boldsymbol{\theta}) = \int_{\mathcal{R}^p} q(\boldsymbol{\phi}, \boldsymbol{\theta})\alpha(\boldsymbol{\phi}, \boldsymbol{\theta})d\boldsymbol{\theta} + \int_{\mathcal{R}^p} r(\boldsymbol{\phi})\delta_{\boldsymbol{\phi}}(d\boldsymbol{\theta}) = (1 - r(\boldsymbol{\phi})) + r(\boldsymbol{\phi}) = 1.$$

Next let us check why $\pi(\boldsymbol{\theta})$ must be the invariant distribution of the transition kernel (11). Note that (11) is same as (8) with $p(\boldsymbol{\phi}, \boldsymbol{\theta}) = q(\boldsymbol{\phi}, \boldsymbol{\theta})\alpha(\boldsymbol{\phi}, \boldsymbol{\theta})$. Thus now if we can show that this $p(\boldsymbol{\phi}, \boldsymbol{\theta})$ satisfies the reversibility condition (9) then following the arguments given in the second paragraph of this sub-section in page 12, $\pi(\boldsymbol{\theta})$ must be the invariant distribution of the transition kernel (11). Thus,

$$
\begin{aligned}
&\pi(\boldsymbol{\phi})p(\boldsymbol{\phi},\boldsymbol{\theta}) \\
&\quad= \pi(\boldsymbol{\phi})q(\boldsymbol{\phi},\boldsymbol{\theta})\alpha(\boldsymbol{\phi},\boldsymbol{\theta}) \\
&\quad= \begin{cases} \pi(\boldsymbol{\phi})q(\boldsymbol{\phi},\boldsymbol{\theta})\frac{\pi(\boldsymbol{\theta})q(\boldsymbol{\theta},\boldsymbol{\phi})}{\pi(\boldsymbol{\phi})q(\boldsymbol{\phi},\boldsymbol{\theta})} & \text{if } \pi(\boldsymbol{\phi})q(\boldsymbol{\phi},\boldsymbol{\theta}) > \pi(\boldsymbol{\theta})q(\boldsymbol{\theta},\boldsymbol{\phi}) \\ \pi(\boldsymbol{\phi})q(\boldsymbol{\phi},\boldsymbol{\theta}) & \text{if } \pi(\boldsymbol{\phi})q(\boldsymbol{\phi},\boldsymbol{\theta}) \le \pi(\boldsymbol{\theta})q(\boldsymbol{\theta},\boldsymbol{\phi}) \end{cases} \\
&\quad= \begin{cases} \pi(\boldsymbol{\theta})q(\boldsymbol{\theta},\boldsymbol{\phi}) & \text{if } \pi(\boldsymbol{\phi})q(\boldsymbol{\phi},\boldsymbol{\theta}) > \pi(\boldsymbol{\theta})q(\boldsymbol{\theta},\boldsymbol{\phi}) \\ \pi(\boldsymbol{\theta})q(\boldsymbol{\theta},\boldsymbol{\phi})\frac{\pi(\boldsymbol{\phi})q(\boldsymbol{\phi},\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})q(\boldsymbol{\theta},\boldsymbol{\phi})} & \text{if } \pi(\boldsymbol{\phi})q(\boldsymbol{\phi},\boldsymbol{\theta}) \le \pi(\boldsymbol{\theta})q(\boldsymbol{\theta},\boldsymbol{\phi}) \end{cases} \\
&\quad= \pi(\boldsymbol{\theta})q(\boldsymbol{\theta},\boldsymbol{\phi})\text{Minimum}\left\{\frac{\pi(\boldsymbol{\phi})q(\boldsymbol{\phi},\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})q(\boldsymbol{\theta},\boldsymbol{\phi})}, 1\right\} \\
&\quad= \pi(\boldsymbol{\theta})q(\boldsymbol{\theta},\boldsymbol{\phi})\alpha(\boldsymbol{\theta},\boldsymbol{\phi}) \\
&\quad= \pi(\boldsymbol{\theta})p(\boldsymbol{\theta},\boldsymbol{\phi})
\end{aligned}
$$

showing that $p(\cdot,\cdot)$ satisfies the reversibility condition (9).

Thus according to the probabilistic interpretation of the transition kernel given in the discussion following (11), the Metropolis-Hastings algorithm may be summarized as follows. Start with an arbitrary initial value $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. In the n-th iteration suppose the current value of $\boldsymbol{\theta}$ is $\boldsymbol{\theta}_n$. Then first generate a candidate value $\boldsymbol{\phi}$ from a proposal density $q(\boldsymbol{\theta}_n,\cdot)$. If $\alpha(\boldsymbol{\theta}_n,\boldsymbol{\phi}) = 1$ set $\boldsymbol{\theta}_{n+1} = \boldsymbol{\phi}$ and proceed to the $(n+1)$-st iteration. If $\alpha(\boldsymbol{\theta}_n,\boldsymbol{\phi}) < 1$ then generate an $\upsilon \sim$Uniform$[0,1]$, and set $\boldsymbol{\theta}_{n+1} = \boldsymbol{\phi}$ if $\upsilon < \alpha(\boldsymbol{\theta}_n,\boldsymbol{\phi})$ otherwise set $\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n$, and in either case proceed to the $(n+1)$-st iteration. Note that in this algorithm one needs the target density $\pi(\boldsymbol{\theta})$ only for the computation of $\alpha(\cdot,\cdot)$, and from the definition of $\alpha(\boldsymbol{\phi},\boldsymbol{\theta})$, it is clear that since it depends on $\pi(\cdot)$ only through the ratio $\pi(\boldsymbol{\theta})/\pi(\boldsymbol{\phi})$, like the rejection method in §2.4, this algorithm also does not require a knowledge about the normalizing constant of $\pi(\cdot)$, and thus making the Metropolis-Hastings algorithm a very attractive choice as a method for generating observations from a posterior distribution specified through (1).

4.2 Gibbs Sampling

By far the most popular method of generating observations from a multi-dimensional posterior $\pi(\boldsymbol{\theta})$ has been the so-called Gibbs sampling. Though it may be formulated as a special case of the Metropolis-Hastings algorithm using the product of kernel property, here we shall take a more direct approach. By this we mean, we shall first describe the algorithm, show that this yields a transition density of an MC, and then show that the target density $\pi(\boldsymbol{\theta})$ is the invariant distribution of this transition density.

Suppose $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_j,\ldots,\theta_p) \in \mathcal{R}^p$. In the sequel we have to deal with all the p one-dimensional conditional distributions of each component given the rest. To simplify notation, we shall generically use $\pi(\cdot|\cdots)$ for the conditional density of the component which does not appear in the subscripts of the conditioning variables. Thus for example $\pi(\theta_j|\theta_1,\ldots,\theta_{j-1},\phi_{j+1},\ldots,\phi_p)$ will denote the conditional density of the j-th component given the remaining $(p-1)$ fixed at their respective values $\theta_1,\ldots,\theta_{j-1},\phi_{j+1},\ldots,\phi_p$. Similarly joint density of a subset of the components will be identified by the subscripts of the dummy variables appearing in $\pi(\cdots)$ *i.e.* for example, $\pi(\theta_1, \ldots, \theta_{j-1}, \phi_{j+1}, \ldots, \phi_p)$ will denote the joint density of the $(p-1)$ components consisting of all but the j-th component of the original p.

The algorithm provides a transition $\boldsymbol{\phi} \to \boldsymbol{\theta}$ as follows. Given an initial value $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)$, first generate a θ_1 from $\pi(\cdot|\phi_2, \ldots, \phi_p)$, the conditional density of the first component given the remaining $(p-1)$ components fixed at their initial values. Then generate a θ_2 from $\pi(\cdot|\theta_1, \phi_3, \ldots, \phi_p)$, the conditional density of the second component after fixing the first component at its just generated value θ_1 and the remaining $(p-2)$ components fixed at their initial values. In general for $1 < j < p$, generate a θ_j from $\pi(\cdot|\theta_1, \ldots, \theta_{j-1}, \phi_{j+1}, \ldots, \phi_p)$, the conditional density of the j-th component after fixing the values of the first $(j-1)$ components at their generated values $\theta_1, \ldots, \theta_{j-1}$ and the remaining $(p-j)$ components fixed at their initial values $(\phi_{j+1}, \ldots, \phi_p)$. And finally generate θ_p from $\pi(\cdot|\theta_1, \ldots, \theta_{p-1})$, the conditional density of the p-th component given the remaining $(p-1)$ components fixed at their generated values. This gives us a transition from $\boldsymbol{\phi} \to \boldsymbol{\theta}$. The steps involved in this transition is pictorially depicted in the following diagram, where the density used for generating the univariate random variable at each stage is indicated above the arrow, and the resulting current state of the p-tuple after the arrow.

$$(\phi_1, \ldots, \phi_p) \xrightarrow{\pi(\theta_1|\phi_2,\ldots,\phi_p)} (\theta_1, \phi_2, \ldots, \phi_p) \xrightarrow{\pi(\theta_2|\theta_1,\phi_3,\ldots,\phi_p)} (\theta_1, \theta_2, \phi_3, \ldots, \phi_p) \longrightarrow \cdots$$

$$(\theta_1, \ldots, \theta_{j-1}, \phi_j, \ldots, \phi_p) \xrightarrow{\pi(\theta_j|\theta_1,\ldots,\theta_{j-1},\phi_{j+1},\ldots,\phi_p)} (\theta_1, \ldots, \theta_j, \phi_{j+1}, \ldots, \phi_p) \longrightarrow \cdots$$

$$(\theta_1, \ldots, \theta_{p-2}, \phi_{p-1}, \phi_p) \xrightarrow{\pi(\theta_{p-1}|\theta_1,\ldots,\theta_{p-2},\phi_p)} (\theta_1, \ldots, \theta_{p-1}, \phi_p) \xrightarrow{\pi(\theta_p|\theta_1,\ldots,\theta_{p-1})} (\theta_1, \ldots, \theta_p)$$

It is easy to see that the above transition $\boldsymbol{\phi} \to \boldsymbol{\theta}$ has the transition density $p(\boldsymbol{\phi}, \boldsymbol{\theta})$ given by,

$$p(\boldsymbol{\phi}, \boldsymbol{\theta}) = \pi(\theta_1|\phi_2, \ldots, \phi_p) \left\{ \prod_{j=2}^{p-1} \pi(\theta_j|\theta_1, \ldots, \theta_{j-1}, \phi_{j+1}, \ldots, \phi_p) \right\} \pi(\theta_p|\theta_1, \ldots, \theta_{p-1}). \tag{12}$$

Note that $p(\boldsymbol{\phi}, \boldsymbol{\theta})$ is a density because being a product of densities it is non-negative and

$$\begin{aligned}
&\int_{\mathcal{R}^p} p(\boldsymbol{\phi}, \boldsymbol{\theta}) d\boldsymbol{\theta} \\
&= \int_{\mathcal{R}^{p-1}} \left[\pi(\theta_1|\phi_2, \ldots, \phi_p) \left\{ \prod_{j=2}^{p-1} \pi(\theta_j|\theta_1, \ldots, \theta_{j-1}, \phi_{j+1}, \ldots, \phi_p) \right\} \left\{ \int_{\mathcal{R}} \pi(\theta_p|\theta_1, \ldots, \theta_{p-1}) d\theta_p \right\} \right] \\
&\qquad \prod_{j=1}^{p-1} d\theta_j \\
&= \int_{\mathcal{R}^{p-2}} \left[\pi(\theta_1|\phi_2, \ldots, \phi_p) \left\{ \prod_{j=2}^{p-2} \pi(\theta_j|\theta_1, \ldots, \theta_{j-1}, \phi_{j+1}, \ldots, \phi_p) \right\} \right. \\
&\qquad \left. \left\{ \int_{\mathcal{R}} \pi(\theta_{p-1}|\theta_1, \ldots, \theta_{p-2}, \phi_p) d\theta_{p-1} \right\} \right] \prod_{j=1}^{p-2} d\theta_j
\end{aligned}$$

....

$$
\begin{aligned}
&= \int_{\mathcal{R}} \left[\pi(\theta_1|\phi_2,\ldots,\phi_p)\left\{\int_{\mathcal{R}} \pi(\theta_2|\theta_1,\phi_3,\ldots,\phi_p)d\theta_2\right\}\right] d\theta_1 \\
&= \int_{\mathcal{R}} \pi(\theta_1|\phi_2,\ldots,\phi_p)d\theta_1 \\
&= 1.
\end{aligned}
$$

Thus $p(\boldsymbol{\phi},\boldsymbol{\theta})$ given in (12) above indeed defines a transition density for the transition $\boldsymbol{\phi} \rightarrow \boldsymbol{\theta}$.

Now let us see why the target density $\pi(\boldsymbol{\theta})$ is the invariant distribution for the transition density given in (12). In order to show this we now appeal to the definition of the invariant distribution for the density case given in (4). Thus starting with the r.h.s. of (4) with $p(\boldsymbol{\phi},\boldsymbol{\theta})$ given in (12) we get,

$$
\begin{aligned}
&\int_{\mathcal{R}^p} p(\boldsymbol{\phi},\boldsymbol{\theta})\pi(\boldsymbol{\phi})d\boldsymbol{\phi} \\
&= \pi(\theta_p|\theta_1,\ldots,\theta_{p-1})\int_{\mathcal{R}^{p-1}}\left[\left\{\prod_{j=2}^{p-1}\pi(\theta_j|\theta_1,\ldots,\theta_{j-1},\phi_{j+1},\ldots,\phi_p)\right\}\right. \\
&\qquad\qquad \left.\pi(\theta_1|\phi_2,\ldots,\phi_p)\pi(\phi_2,\ldots,\phi_p)\right]\prod_{j=2}^{p} d\phi_j \\
&= \pi(\theta_p|\theta_1,\ldots,\theta_{p-1})\int_{\mathcal{R}^{p-2}}\left[\left\{\prod_{j=2}^{p-1}\pi(\theta_j|\theta_1,\ldots,\theta_{j-1},\phi_{j+1},\ldots,\phi_p)\right\}\right. \\
&\qquad\qquad \left.\int_{\mathcal{R}}\pi(\theta_1,\phi_2,\ldots,\phi_p)d\phi_2\right]\prod_{j=3}^{p} d\phi_j \\
&= \pi(\theta_p|\theta_1,\ldots,\theta_{p-1})\int_{\mathcal{R}^{p-2}}\left[\left\{\prod_{j=2}^{p-1}\pi(\theta_j|\theta_1,\ldots,\theta_{j-1},\phi_{j+1},\ldots,\phi_p)\right\}\pi(\theta_1,\phi_3,\ldots,\phi_p)\right]\prod_{j=3}^{p} d\phi_j \\
&= \pi(\theta_p|\theta_1,\ldots,\theta_{p-1})\int_{\mathcal{R}^{p-3}}\left[\left\{\prod_{j=3}^{p-1}\pi(\theta_j|\theta_1,\ldots,\theta_{j-1},\phi_{j+1},\ldots,\phi_p)\right\}\right. \\
&\qquad\qquad \left.\int_{\mathcal{R}}\pi(\theta_2|\theta_1,\phi_3,\ldots,\phi_p)\pi(\theta_1,\phi_3,\ldots,\phi_p)d\phi_3\right]\prod_{j=4}^{p} d\phi_j \\
&= \pi(\theta_p|\theta_1,\ldots,\theta_{p-1})\int_{\mathcal{R}^{p-3}}\left[\left\{\prod_{j=3}^{p-1}\pi(\theta_j|\theta_1,\ldots,\theta_{j-1},\phi_{j+1},\ldots,\phi_p)\right\}\pi(\theta_1,\theta_2,\phi_4,\ldots,\phi_p)\right]\prod_{j=4}^{p} d\phi_j \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
&= \pi(\theta_p|\theta_1,\ldots,\theta_{p-1})\int_{\mathcal{R}^{p-k}}\left[\left\{\prod_{j=k}^{p-1}\pi(\theta_j|\theta_1,\ldots,\theta_{j-1},\phi_{j+1},\ldots,\phi_p)\right\}\right. \\
&\qquad\qquad \left.\int_{\mathcal{R}}\pi(\theta_{k-1}|\theta_1,\ldots,\theta_{k-2},\phi_k,\ldots,\phi_p)\pi(\theta_1,\ldots,\theta_{k-2},\phi_k,\ldots,\phi_p)d\phi_k\right]\prod_{j=k+1}^{p} d\phi_j \\
&= \pi(\theta_p|\theta_1,\ldots,\theta_{p-1})\int_{\mathcal{R}^{p-k}}\left[\left\{\prod_{j=k}^{p-1}\pi(\theta_j|\theta_1,\ldots,\theta_{j-1},\phi_{j+1},\ldots,\phi_p)\right\}\right. \\
&\qquad\qquad \left.\pi(\theta_1,\ldots,\theta_{k-1},\phi_{k+1},\ldots,\phi_p)\right]\prod_{j=k+1}^{p} d\phi_j
\end{aligned}
$$

$$
\begin{aligned}
& \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
&= \pi(\theta_p|\theta_1,\ldots,\theta_{p-1}) \int_{\mathcal{R}} \pi(\theta_1,\ldots,\theta_{p-1},\phi_p)d\phi_p \\
&= \pi(\theta_1,\ldots,\theta_p)
\end{aligned}
$$

showing that the target density $\pi(\boldsymbol{\theta})$ is the invariant distribution for the transition density given in (12). Thus a sample from $\pi(\boldsymbol{\theta})$ may be drawn as follows. Start with an arbitrary initial value $\boldsymbol{\theta}_0$. For the n-th transition from $\boldsymbol{\theta}_n \to \boldsymbol{\theta}_{n+1}$, follow the transition diagram given in page 15 with ϕ_j replaced by θ_{jn} and θ_j replaced by $\theta_{j(n+1)}$. For large n, $\boldsymbol{\theta}_n$ may be construed as a sample from the joint target density $\pi(\boldsymbol{\theta})$.

5 Examples

In this section we provide a few examples of application of MCMC techniques for Bayesian analysis of some standard statistical models.

5.1 Logistic Regression

Consider the logistic regression model for a 0-1 valued response variable Y, given a set of k covariates $\boldsymbol{X} = (X_1,\ldots,X_k) = (x_1,\ldots,x_k) = \boldsymbol{x}$ with $P(Y=1|\boldsymbol{X}=\boldsymbol{x})$ modeled as

$$
P(Y=1|\boldsymbol{X}=\boldsymbol{x}) = \frac{1}{1+\exp\left(-\beta_0-\beta_1 x_1-\cdots-\beta_k x_k\right)} \tag{13}
$$

Given the data set $\boldsymbol{D} = \{(y_1,\boldsymbol{x}_1),\ldots,(y_n,\boldsymbol{x}_n)\}$ consisting of n i.i.d. observations on Y together with their respective accompanying covariates, we are to draw inference about the $p = (k+1)\times 1$ parameter vector $\boldsymbol{\beta} = (\beta_0,\beta_1,\ldots,\beta_k)'$. According to the logistic regression model (13), the likelihood of $\boldsymbol{\beta}$ is given by

$$
\begin{aligned}
& L(\boldsymbol{\beta}|\boldsymbol{D}) = \qquad\qquad (14) \\
& \prod_{i=1}^{n}\left[\left\{\frac{1}{1+\exp\left(-\beta_0-\beta_1 x_{i1}-\cdots-\beta_k x_{ik}\right)}\right\}^{y_i}\left\{\frac{\exp\left(-\beta_0-\beta_1 x_{i1}-\cdots-\beta_k x_{ik}\right)}{1+\exp\left(-\beta_0-\beta_1 x_{i1}-\cdots-\beta_k x_{ik}\right)}\right\}^{1-y_i}\right]
\end{aligned}
$$

where $\boldsymbol{x}_i = (x_{i1},\ldots,x_{ik})$. Now let us assume that the prior for $\boldsymbol{\beta}$ is p-variate Normal with mean $\boldsymbol{\mu}$ and dispersion $\boldsymbol{\Sigma}$ so that

$$
\pi(\boldsymbol{\beta}) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu})\right\}. \tag{15}
$$

Thus according to (1), $\pi(\boldsymbol{\beta}|\boldsymbol{D})$, the joint posterior of $\boldsymbol{\beta}$, is proportional to the product of (14) and (15), which cannot be recognized as any known multivariate density. Thus the only way to tackle this posterior would be either numerical integration or some way of drawing sample from it. We shall obviously take the second approach with the first method bordering on computational in-feasibility. We shall use Gibbs sampling for drawing sample from this posterior.

In order to apply Gibbs sampling we shall need the conditional posterior of each β_j given the rest, denoted by $\pi(\beta_j|\boldsymbol{\beta}_{(-j)}, \boldsymbol{D})$, for $j = 0, 1, \ldots, k$. Note that the functional form of $\pi(\beta_j|\boldsymbol{\beta}_{(-j)}, \boldsymbol{D})$ is exactly same as the the product of (14) and (15), viewed as a function of β_j with all other variables treated as known constants[1]. After studying (14) and (15), and defining $x_{i0} = 1\ \forall i = 1, \ldots, n$, for $j = 0, 1, \ldots, k$ this form may be abstracted as

$$\pi(\beta_j|\boldsymbol{\beta}_{(-j)}, \boldsymbol{D}) \propto \frac{\exp\left(a_j + b_j\beta_j - c_j^2\beta_j^2\right)}{\prod_{i=1}^n \left\{1 + \exp\left(a_{ij} + b_{ij}\beta_j\right)\right\}} \tag{16}$$

for some constants a_j, b_j, c_j, a_{ij}'s and b_{ij}'s which depend on $\boldsymbol{\beta}_{(-j)}$, $\boldsymbol{D}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. The denominator of (16) comes solely from (14), with $a_{ij} = -\left(\beta_0 + \sum_{\substack{l=1\\l\neq j}}^k \beta_l x_{il}\right)$ and $b_{ij} = -x_{ij}$; while the quadratic term involving β_j^2 in the numerator comes solely from the prior (15), with $c_j^2 = \sigma^{jj}/2$, where $\boldsymbol{\Sigma}^{-1} = ((\sigma^{ij}))$, and $a_j = -\left\{\sum_{i=1}^n (1 - y_i)\left(\beta_0 + \sum_{\substack{l=1\\l\neq j}}^k \beta_l x_{il}\right)\right\} +$ some terms coming in from (15) and $b_j = -\sum_{i=1}^n (1 - y_i)x_{ij} +$ some terms coming in from (15). Now even the expression in the r.h.s. of (16) cannot be recognized in the form of some known univariate density, so that we can immediately start drawing sample from it. However there is a simple technique of drawing samples from densities which are log-concave *i.e.* log of the density is a concave function, which has been explained in the Appendix. Thus if we can show that $\log \pi(\beta_j|\boldsymbol{\beta}_{(-j)}, \boldsymbol{D})$ is a concave function of β_j then we can appeal to this method described in the Appendix to draw samples from $\pi(\beta_j|\boldsymbol{\beta}_{(-j)}, \boldsymbol{D})$. The simplest way to show that a function is concave, is to show that its second derivative is negative. Thus

$$\begin{aligned}
&\frac{\partial^2}{\partial\beta_j^2}\left[\log \pi(\beta_j|\boldsymbol{\beta}_{(-j)}, \boldsymbol{D})\right] \\
&= \frac{\partial^2}{\partial\beta_j^2}\left[\left(a_j + b_j\beta_j - c_j^2\beta_j^2\right) - \sum_{i=1}^n \log\left\{1 + \exp\left(a_{ij} + b_{ij}\beta_j\right)\right\}\right] \\
&= \frac{\partial}{\partial\beta_j}\left[\left(b_j - 2c_j^2\beta_j\right) - \sum_{i=1}^n \frac{b_{ij}\exp\left(a_{ij} + b_{ij}\beta_j\right)}{1 + \exp\left(a_{ij} + b_{ij}\beta_j\right)}\right] \\
&= -2c_j^2 - \sum_{i=1}^n \frac{b_{ij}^2 \exp\left(a_{ij} + b_{ij}\beta_j\right)}{\left\{1 + \exp\left(a_{ij} + b_{ij}\beta_j\right)\right\}^2} \\
&< 0
\end{aligned}$$

establishing that $\pi(\beta_j|\boldsymbol{\beta}_{(-j)}, \boldsymbol{D})$ is log-concave. Thus one can now draw samples from (16) using its log-concavity and the algorithm described in the Appendix for each j. Now since we have a way to generate samples from the conditional posteriors of each component given the rest, we can proceed with the Gibbs sampling as described in §4.2 towards generating samples from $\pi(\boldsymbol{\beta}|\boldsymbol{D})$, the joint posterior of $\boldsymbol{\beta}$.

[1] This observation has nothing to do with the logistic regression example at hand. It is true for any multi-dimensional density $\pi(\theta_1, \ldots, \theta_p)$. Functional form of the conditional density of any θ_j given $\boldsymbol{\theta}_{(-j)} = (\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_p)$ is same as $\pi(\theta_1, \ldots, \theta_p)$, viewed as a function of only θ_j with the remaining $(p-1)$ θ_l's treated as known constants. This feature is one of the major sources of popularity of Gibbs sampling.

5.2 Normal Mixtures

Suppose the random variable Y has density

$$f(y|\boldsymbol{\mu},\boldsymbol{\tau},\boldsymbol{\pi}) = \sum_{j=1}^{J} \pi_j \frac{\tau_j^{1/2}}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\tau_j(y-\mu_j)^2\right\} \tag{17}$$

where $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_J)$ and $\boldsymbol{\tau} = (\tau_1,\ldots,\tau_J)$ are respective mean and precision parameters of J Normal distributions, and $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_J)$ is the mixture probability vector with $0 < \pi_j < 1$ and $\sum_{j=1}^{J}\pi_j = 1$. The density in (17) is called a mixture of Normal distributions with mixing proportion $\boldsymbol{\pi}$. Given an i.i.d. sample $Y_1, Y_2, \ldots, Y_n$ on Y we are to draw inference on $(\boldsymbol{\mu},\boldsymbol{\tau},\boldsymbol{\pi})$. The first step towards the computation of posterior is the calculation of the likelihood of $(\boldsymbol{\mu},\boldsymbol{\tau},\boldsymbol{\pi})$ given the data $\boldsymbol{Y} = \boldsymbol{y}$, which is given by

$$L(\boldsymbol{\mu},\boldsymbol{\tau},\boldsymbol{\pi}|\boldsymbol{y}) = \prod_{i=1}^{n}\left[\sum_{j=1}^{J} \pi_j \frac{\tau_j^{1/2}}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\tau_j(y_i-\mu_j)^2\right\}\right]. \tag{18}$$

Now, as is customary with Normal parameters, let us put J independent Normal-Gamma priors with parameters $(\theta_j, \psi_j, \beta_j, \lambda_j)$ on (μ_j, τ_j) so that

$$\pi(\boldsymbol{\mu},\boldsymbol{\tau}) = \prod_{j=1}^{J}\left[\frac{\lambda_j^{\beta_j}\psi_j^{1/2}}{\Gamma(\beta_j)\sqrt{2\pi}}\tau_j^{\beta_j-1/2} \exp\left\{-\frac{1}{2}\tau_j\left[2\lambda_j + \psi_j\left(\mu_j-\theta_j\right)^2\right]\right\}\right], \tag{19}$$

and independently put a Dirichlet prior with parameters $(\alpha_1,\ldots,\alpha_J)$ on the mixing proportion parameters $\boldsymbol{\pi}$, which automatically satisfies the constraints $0 < \pi_j < 1$ and $\sum_{j=1}^{J}\pi_j = 1$ so that

$$\pi(\boldsymbol{\pi}) = \frac{\Gamma\left(\sum_{j=1}^{J}\alpha_j\right)}{\prod_{j=1}^{J}\Gamma(\alpha_j)}\prod_{j=1}^{J}\pi_j^{\alpha_j-1}. \tag{20}$$

By (1), $\pi(\boldsymbol{\mu},\boldsymbol{\tau},\boldsymbol{\pi}|\boldsymbol{y})$, the posterior of $(\boldsymbol{\mu},\boldsymbol{\tau},\boldsymbol{\pi})$, is proportional to the product of (18), (19) and (20), which is a very messy expression. But this problem is circumvented as follows. Introduce a random variable I, which takes values in $\{1,\ldots,J\}$ such that its marginal p.m.f. is given by $\boldsymbol{\pi}$ *i.e.* $P(I=j) = \pi_j$ for $j = 1,\ldots,J$, and conditionally $Y|I \sim N(\mu_I, 1/\tau_I^2)$. Then it is easy to see that, according to these marginal of I and conditional of $Y|I$, the marginal density of Y is same as that is given in (17). In order to understand the physical significance of this newly introduced random variable I, let us try to understand how observations are generated from a mixture distribution such as the one given in (17), which is a mixture of J Normal populations. In order to generate such an Y, we first select the index of the candidate population, from which the final observation will be drawn, from the available J, generated according to the p.m.f. $\boldsymbol{\pi}$. Once this index is generated, an observation is then drawn from this selected population, which in this case happens to be Normal. Thus this index random variable I, though might remain unobserved, lies at the very heart of the definition of a mixture distribution. The I thus defined is called a latent variable or auxiliary variable or pseudo-data, as like a parameter it remains unobserved or latent along with the observed data $\boldsymbol{Y} = \boldsymbol{y}$. But now if we augment each Y_i with its corresponding I_i, so that our new data set becomes $(\boldsymbol{Y}, \boldsymbol{I})$, then conditional on the parameters $(\boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\pi})$, the joint distribution of $(\boldsymbol{Y}, \boldsymbol{I})$ is given by

$$f(\boldsymbol{y}, \boldsymbol{i} | \boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\pi}) = \prod_{k=1}^{n} \left[\pi_{i_k} \frac{\tau_{i_k}^{1/2}}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} \tau_{i_k} (y_k - \mu_{i_k})^2 \right\} \right] \tag{21}$$

which is same as the likelihood of $(\boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\pi})$ given $(\boldsymbol{Y}, \boldsymbol{I}) = (\boldsymbol{y}, \boldsymbol{i})$.

But note that we are to find the posterior of $(\boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\pi})$ given $\boldsymbol{Y} = \boldsymbol{y}$ and not $(\boldsymbol{Y}, \boldsymbol{I}) = (\boldsymbol{y}, \boldsymbol{i})$. In order to do so, we shall consider $\pi(\boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\pi}, \boldsymbol{i} | \boldsymbol{y})$, the conditional joint distribution of $(\boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\pi}, \boldsymbol{I})$ given $\boldsymbol{Y} = \boldsymbol{y}$; generate observations from $\pi(\boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\pi}, \boldsymbol{I} | \boldsymbol{y})$; and then look at the generated values of $(\boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\pi})$ from this distribution, which in particular will constitute a sample from $\pi(\boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\pi} | \boldsymbol{y})$, the posterior of $(\boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\pi})$ given $\boldsymbol{Y} = \boldsymbol{y}$. But generating observations from $\pi(\boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\pi}, \boldsymbol{i} | \boldsymbol{y})$ is now easy, as it is proportional to the product of (19), (20) and (21) viewed as a function of $(\boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\pi}, \boldsymbol{i})$ with $\boldsymbol{y}$ as fixed constants. By studying this product we can immediately propose a Gibbs sampling scheme for generating $(\boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\pi}, \boldsymbol{i})$ from $\pi(\boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\pi}, \boldsymbol{i} | \boldsymbol{y})$ as follows.

Step 1: Given $(\boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\pi}, \boldsymbol{i}_{(-k)}, \boldsymbol{y})$, where $\boldsymbol{i}_{(-k)} = (i_1, \ldots, i_{k-1}, i_{k+1}, \ldots, i_n)$ are the $(n-1)$ remaining generated values of $\boldsymbol{I}$, generate an I_k from $\{1, \ldots, J\}$ according to the p.m.f. $\boldsymbol{\pi}$, for $k = 1, \ldots, n$.

Step 2: Given $(\boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{i}, \boldsymbol{y})$, generate $\boldsymbol{\pi}$ from a Dirichlet distribution with parameters $(\alpha_1 + \sum_{k=1}^{n} \chi_{[i_k=1]}, \ldots, \alpha_j + \sum_{k=1}^{n} \chi_{[i_k=j]}, \ldots, \alpha_j + \sum_{k=1}^{n} \chi_{[i_k=J]})$, where $\underset{A}{\chi} = \begin{cases} 1 & \text{if } A \text{ is true} \\ 0 & \text{otherwise} \end{cases}$.

Step 3: Given $(\boldsymbol{\mu}_{(-j)}, \boldsymbol{\tau}, \boldsymbol{\pi}, \boldsymbol{i}, \boldsymbol{y})$, where $\boldsymbol{\mu}_{(-j)} = (\mu_1, \ldots, \mu_{j-1}, \mu_{j+1}, \ldots, \mu_J)$, generate μ_j from a Normal distribution with mean $\frac{n_j \overline{y}_j + \psi_j \theta_j}{n_j + \psi_j}$ and precision $(n_j + \psi_j)\tau_j$, where $n_j = \sum_{k=1}^{n} \chi_{[i_k=j]}$ and $\overline{y}_j = \frac{1}{n_j} \sum_{k=1}^{n} y_k \chi_{[i_k=j]}$.

Step 4: Given $(\boldsymbol{\mu}, \boldsymbol{\tau}_{(-j)}, \boldsymbol{\pi}, \boldsymbol{i}, \boldsymbol{y})$, where $\boldsymbol{\tau}_{(-j)} = (\tau_1, \ldots, \tau_{j-1}, \tau_{j+1}, \ldots, \tau_J)$, generate τ_j from a Gamma distribution with shape parameter $(n_j+1)/2+\beta_j$ and scale parameter $(\lambda_j + s_j^2/2) + \frac{n_j+\psi_j}{2} \left(\mu_j - \frac{n_j \overline{y}_j + \psi_j \theta_j}{n_j + \psi_j} \right)^2$, where $s_j^2 = \frac{1}{n_j} \sum_{k=1}^{n} (y_k - \overline{y}_j)^2 x 1 \chi_{[i_k=j]}$.

The four steps detailed above provide a transition from $(\boldsymbol{\mu}_n, \boldsymbol{\tau}_n, \boldsymbol{\pi}_n, \boldsymbol{i}_n) \rightarrow (\boldsymbol{\mu}_{n+1}, \boldsymbol{\tau}_{n+1}, \boldsymbol{\pi}_{n+1}, \boldsymbol{i}_{n+1})$ in one step of a Gibbs iteration expressed in terms of the required conditional distributions. Thus a sample from $\pi(\boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\pi}, \boldsymbol{i} | \boldsymbol{y})$ may be obtained by iterating over these Gibbs iterations a large number of times and then choosing the final value.

Appendix: Log-Concave Densities

Many a times, as in the Logistic Regression example in §5.1, in the Gibbs or Metropolis-Hastings sample generation schemes, we encounter densities which do not appear to belong to any known family of probability densities so that we can readily start sampling from them using known algorithms *e.g.* (16). In such situations however as in the case of the Logistic Regression example in §5.1, sometimes we might be able to show that the density at hand is log-concave. Now if a univariate density is log-concave, then we can employ a simple rejection algorithm discussed in §2.4 to obtain a sample from such densities, without requiring any knowledge about the normalizing constant of the density. The purpose of this appendix is to outline this method of obtaining samples from a log-concave univariate density known sans the normalizing constant.

Suppose $\pi(\theta)$ be a density with $\pi(\theta) = c_f f(\theta)$ where c_f is unknown and $f(\theta)$ is known with $\int_{-\infty}^{\infty} f(\theta)\, d\theta = 1/c_f$ so that $\int_{-\infty}^{\infty} \pi(\theta\, d\theta) = 1$. Suppose further that $h(\theta) = \log(f(\theta))$ is concave or its first derivative $h'(\theta)$ is a decreasing function of θ. Start with a grid of k points $\{\theta_1, \ldots, \theta_k\}$ such that $h'(\theta_1) > 0$ and $h'(\theta_k) < 0$. Consider the tangent $t_j(\theta)$ of $h(\theta)$ at each θ_j for $j = 1, \ldots, k$ given by

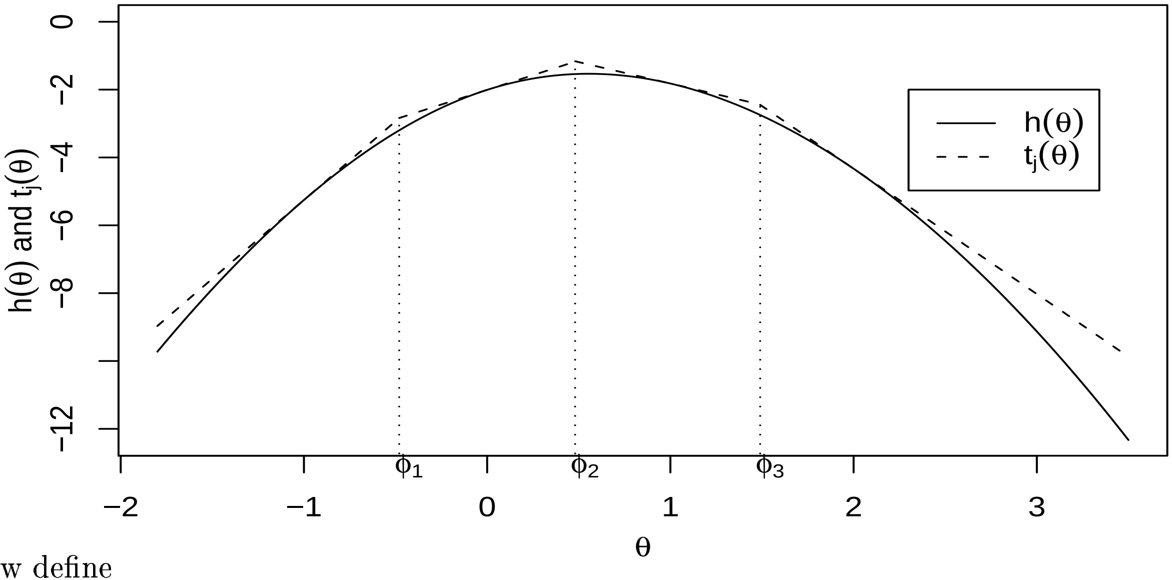
$$t_j(\theta) = h(\theta_j) + h'(\theta_j)(\theta - \theta_j).$$

Now for $j = 1, \ldots, k-1$ consider the abscissa of the point of intersection of $t_j(\theta)$ and $t_{j+1}(\theta)$ given by

$$\phi_j = \frac{(h(\theta_j) - h(\theta_{j+1})) + (\theta_{j+1} h'(\theta_{j+1}) - \theta_j h'(\theta_j))}{h'(\theta_{j+1}) - h'(\theta_j)}.$$

By concavity of $h(\theta)$, $h(\theta) \leq t_j(\theta)\ \forall \theta \in [\phi_{j-1}, \phi_j]\ \forall j = 1, \ldots, k$ with $\phi_0 = -\infty$ and $\phi_k = +\infty$. This situation is depicted in Figure 2 below.

Figure 2: Envelope Function Log−Concave Densities



Now define

$$g(\theta) = \frac{\exp(t_j(\theta))}{\sum_{i=1}^{k} \int_{\phi_{i-1}}^{\phi_i} \exp(t_i(\phi)) d\phi} \text{ for } \theta \in [\phi_{j-1}, \phi_j].$$

Then $g(\theta)$ is a density which being piecewise exponential, is easy to sample from and $\forall \theta \in (-\infty . \infty)$, $f(\theta) \leq c g(\theta)$ where $c = \sum_{j=1}^{k} \int_{\phi_{j-1}}^{\phi_j} \exp(t_j(\phi)) d\phi$. Thus now sampling from $\pi(\theta)$, the original density of interest, can now proceed with the rejection method described in §2.4.