STAT 8054 notes on Markov Chain Monte Carlo

Spring 2014

May 4, 2014

Contents

1 Introduction

In ordinary Monte Carlo, we studied methods to generate a realization of $X_1, \ldots, X_n$, which are iid with some distribution of interest. Our goal is to still sample from this distribution, but we are willing to relax the requirement that $X_1, \ldots, X_n$ are iid: we will allow $X_1, \ldots, X_n$ to be dependent and have different distributions, but the distribution of X_k will converge to the distribution of interest as k increases.

2 Review of $\mathcal{X}$-valued random variables

The following review is based on Chapter 2 of Keener (2005). Suppose that $(\Omega, \mathcal{F}, P)$ is a probability space and $(\mathcal{X}, \mathcal{B})$ is a measurable space. An $\mathcal{X}$-valued random variable X is function $X : \Omega \to \mathcal{X}$ such that $\{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}$ for every $B \in \mathcal{B}$. The probability measure $P_X : \mathcal{B} \to [0, 1]$ defined by

$$P_X(A) = P(X \in A) \equiv P(\{\omega \in \Omega : X(\omega) \in A\}),$$

is called the *distribution* of X. We indicate this by writing $X \sim P_X$.

Let $h : \mathcal{X} \to \mathbb{R}$ be a measurable function, i.e. $\{x \in \mathcal{X} : h(x) \in B\} \in \mathcal{B}$ for every $B \in \mathcal{B}_{\mathbb{R}}$. We define the integral of h against the probability measure P_X by

$$\int h(x)P_X(dx) = E\{h(X)\}.$$

In particular, when $h(x) = 1(x \in A)$ and $A \in \mathcal{B}$,

$$P_X(A) = \int 1(x \in A)P_X(dx).$$

If ν is a sigma-finite measure on $\mathcal{B}$ such that $P_X(A) = 0$ whenever $\nu(A) = 0$, then we say that P_X is absolutely continuous with respect to ν. In this case, there exits a non-negative measurable function f called the density of P_X with respect to ν such that

$$P_X(A) = \int 1(x \in A)f(x)\nu(dx).$$

Also for $X \sim P_X$ we have that

$$E\{h(X)\} = \int h(x)P_X(dx) = \int h(x)f(x)\nu(dx).$$

Example 1.

Suppose that $(\mathcal{X}, \mathcal{B}) = (\mathbb{R}^p, \mathcal{B}_{\mathbb{R}^p})$, and X is an $\mathbb{R}^p$-valued random variable (typically called a random vector) with distribution P_X with density f with respect to Lebesgue measure ν. Then

$$\begin{aligned} E\{h(X)\} &= \int h(x)P_X(dx) = \int h(x)f(x)\nu(dx) \\ &= \int \cdots \int h(x_1, \ldots, x_p)f(x_1, \ldots, x_p)dx_1, \ldots, dx_p. \end{aligned}$$

Example 2.

Suppose that $\mathcal{X}$ is countable so our measurable space is $(\mathcal{X}, 2^{\mathcal{X}})$. Let X be an $\mathcal{X}$-valued random variable with distribution P_X with density f (called a probability mass function) with respect to counting measure ν. Then

$$E\{h(X)\} = \int h(x)P_X(dx) = \sum_{x \in \mathcal{X}} h(x)f(x).$$

3 Discrete time homogeneous Markov chains

This section is based on Atchade (2008) and Jones (2013).

Definition 1. *A transition kernel* Q *on a measurable space* $(\mathcal{X}, \mathcal{B})$ *is a function* $Q : \mathcal{X} \times \mathcal{B} \to [0, 1]$ *such that* $Q(x, \cdot) : \mathcal{B} \to [0, 1]$ *is a probability measure on* $\mathcal{B}$ *(a distribution) for all* $x \in \mathcal{X}$ *and* $Q(\cdot, A) : \mathcal{X} \to [0, 1]$ *is a measurable function for all* $A \in \mathcal{B}$*.*

Example 3.

Take $\mathcal{X} = \mathbb{R}^p$ and $\mathcal{B} = \mathcal{B}_{\mathbb{R}^p}$. Let ν be Lebesgue measure on $\mathbb{R}^p$ and suppose that $\Sigma \in \mathbb{S}^p_+$. Define

$$Q(x, A) = \int 1(y \in A)(2\pi)^{-p/2}\{\det(\Sigma^{-1})\}^{1/2} \exp\{0.5(y - x)'\Sigma^{-1}(y - x)\}\nu(dy),$$

for all $x \in \mathbb{R}^p$ and $A \in \mathcal{B}$. Then Q is a transition kernel: $Q(x, \cdot)$ is a probability measure on $\mathcal{B}$ (it's the $N_p(x, \Sigma)$ distribution) and $\{x \in \mathbb{R}^p : Q(x, A) \in B\} \in \mathcal{B}_{\mathbb{R}^p}$ for every $B \in \mathcal{B}_{\mathbb{R}}$.

Definition 2. *A sequence of* $\mathcal{X}$*-valued random variables* $X_0, X_1, \ldots,$ *is a (time-homogeneous) Markov chain with transition kernel* Q *if*

$$(X_{n+1}|X_0 = x_0, \ldots, X_n = x_n) \sim (X_{n+1}|X_n = x_n) \sim Q(x_n, \cdot),$$

for each non-negative integer n*. Here* $P(X_{n+1} \in A|X_n = x_n) = Q(x_n, A)$ *for all* $A \in \mathcal{B}$*.*

To generate a realization of a Markov chain where X_0 has *initial distribution* P_{X_0} and the chain has transition kernel Q, we

1. generate a realization x_0 of $X_0 \sim P_{X_0}$
2. generate a realization x_1 of $(X_1|X_0 = x_0) \sim Q(x_0, \cdot)$
3. generate a realization x_2 of $(X_2|X_1 = x_1) \sim Q(x_1, \cdot)$
4. ...

the sequence $x_0, x_1, \ldots$ is a realization of the Markov chain. Let $X_0 \sim P_{X_0}$, then the marginal distribution/probability measure for X_1 at $A \in \mathcal{B}$ is

$$\begin{aligned}
P(X_1 \in A) &= E[E\{1(X_1 \in A)|X_0\}] \\
&= \int \left\{ \int 1(x_1 \in A)Q(x_0, dx_1) \right\} P_{X_0}(dx_0) \\
&= \int Q(x_0, A)P_{X_0}(dx_0) \\
&\equiv P_{X_0}Q(A).
\end{aligned}$$

We write this as $X_1 \sim P_{X_0}Q$. We also have that $(X_2|X_0 = x_0) \sim QQ(x_0, \cdot)$:

$$\begin{aligned}
P(X_2 \in A|X_0 = x_0) &= E[E\{1(X_2 \in A)|X_1, X_0 = x_0\}|X_0 = x_0] \\
&= \int \left\{ \int 1(x_2 \in A) Q(x_1, dx_2) \right\} Q(x_0, dx_1) \\
&= \int Q(x_1, A) Q(x_0, dx_1) \\
&\equiv QQ(x_0, A)
\end{aligned}$$

So $X_2 \sim P_{X_0}QQ = P_{X_0}Q^{(2)}$, In general, $X_n \sim P_{X_0}Q^{(n)}$ and $(X_n|X_0 = x) \sim Q^{(n)}(x, \cdot)$.

Suppose that $h : \mathcal{X} \to \mathbb{R}$ is measurable, i.e. $\{x \in \mathcal{X} : h(x) \in B\} \in \mathcal{B}$ for every Borel set $B \in \mathcal{B}_{\mathbb{R}}$. Then

$$E\{h(X_{n+1})|X_n = x_n\} = \int h(x_{n+1}) Q(x_n, dx_{n+1}) \equiv Qh(x_n),$$

for each $x_n \in \mathcal{X}$.

If $X_0 \sim P_X$ implies that $X_1 \sim P_X$, then P_X is an *invariant distribution* for the transition kernel Q, i.e.

$$P_X(A) = \int Q(x, A) P_X(dx),$$

for all $A \in \mathcal{B}$. We write this as $P_X = P_X Q$. If the initial distribution is P_X, then $X_k \sim P_X$ for all k. We say that Q is *reversible* with respect to P_X if

$$\iint h(x, y) Q(x, dy) P_X(dx) = \iint h(x, y) Q(y, dx) P_X(dy), \tag{1}$$

for any h such that both sides of (1) exist. If the chain is reversible, then the distribution of $(X_i, X_{i+1}, \ldots, X_{i+k})$ is the same as the distribution of $(X_{i+k}, \ldots, X_i)$ for all i and k. Reversible implies invariance:

$$\begin{aligned}
P_X Q(A) &= \int Q(x, A) P_X(dx) \\
&= \iint 1(y \in A) Q(x, dy) P_X(dx) \\
&= \iint 1(y \in A) Q(y, dx) P_X(dy) \\
&= \int 1(y \in A) P_X(dy) \left(\int Q(y, dx) \right) \\
&= P_X(A).
\end{aligned}$$

Suppose that $q(x, \cdot)$ is a density for $Q(x, \cdot)$ and f is a density for P_X. If the *detailed balance* condition holds, i.e.

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

then P_X is invariant for Q.

4 Convergence of Markov Chains used in MCMC

Let $(\mathcal{X}, \mathcal{B})$ be a measurable space and let $X_0, X_1, \ldots$ be an $\mathcal{X}$-valued Markov Chain with transition kernel Q and invariant distribution P_X.

1. The chain is *aperiodic* if there does not exist a $d \geq 2$ and disjoint subsets $\mathcal{X}_1, \ldots, \mathcal{X}_d$ of $\mathcal{X}$ such that $Q(x, \mathcal{X}_1) = 1$ for all $x \in \mathcal{X}_d$ and $Q(x, \mathcal{X}_i) = 1$ for all $x \in \mathcal{X}_{i-1}$ $(i = 2, \ldots, d)$.
2. The chain is P_X-*irreducible* if for any $A \in \mathcal{B}$ with $P_X(A) > 0$ implies that $P(t_A < \infty | X_0 = x) > 0$ for $x \in \mathcal{X}$, where $t_A = \min(n > 0 : X_n \in A)$.
3. The chain is *Harris* if for any $A \in \mathcal{B}$ with $P_X(A) > 0$ implies that $P(t_A < \infty | X_0 = x) = 1$ for all $x \in \mathcal{X}$, where $t_A = \min(n > 0 : X_n \in A)$.

Define the total variation distance between measures P_X and P_Y by

$$\|P_X - P_Y\|_{\mathrm{TV}} = \sup_{A \in \mathcal{B}} |P_X(A) - P_Y(A)|$$

The norm is $\|P_Z\|_{\mathrm{TV}} = \sup_{A \in \mathcal{B}} P_Z(A) - \inf_{A \in \mathcal{B}} P_Z(A)$.

Theorem 1. *If the $\mathcal{X}$-valued Markov Chain $X_0, X_1, \ldots$ with invariant distribution P_X and transition kernel Q is P_X-irreducible, then P_X is the unique invariant distribution. Furthermore, if the chain is aperiodic, then there exists a set N with $P_X(N) = 0$ such that for all $x \notin N$,*

$$\|Q^{(n)}(x, \cdot) - P_X(\cdot)\|_{\mathrm{TV}} \to 0,$$

as $n \to \infty$. This holds for all $x \in \mathcal{X}$ if the chain is Harris. This result is taken from "Convegence of Markov chains from all starting points with applications to Metropolis–Hastings algorithms" by R. L. Tweedie in 1999.

5 Algorithms

5.1 Metropolis–Hastings

Let $(\mathcal{X}, \mathcal{B})$ be a measurable space. Our goal is to approximately sample from the distribution P_X. We will create an $\mathcal{X}$-valued Markov chain $X_0, X_1, \ldots$ with transition kernel Q so that the target distribution P_X is invariant for Q. Let P_X have density f with respect to the sigma-finite measure ν: $P_X(dx) = f(x)\nu(dx)$. Let G be a proposal kernel where $G(x, dy) = g(x, y)\nu(dy)$, so $g(x, \cdot)$ is the density for $G(x, \cdot)$ with respect to ν. Define

$$\alpha(x, y) = \min\left\{1, \frac{f(y)g(y, x)}{f(x)g(x, y)}\right\}. \tag{2}$$

The denominator of the *Hastings ratio* in (2) is always positive provided that we started the chain at a point $x \in \mathcal{X}$ such that $f(x) > 0$. Of course, $g(x, y) > 0$ because y is generated from the distribution with density $g(x, \cdot)$. In (2), we only need to know unnormalized densities $\tilde{f}(x) = K_f f(x)$ and $\tilde{g}(x, y) = K_g g(x, y)$, because these constants cancel in the Hastings ratio.

Algorithm 1. *Pick or generate $X_0 \in \mathcal{X}$ with $f(X_0) > 0$ and set $n = 0$*

1. *Given $X_n = x_n$, generate $Z \sim G(x_n, \cdot)$ and independently generate $U \sim \mathrm{Unif}(0,1)$.*
2. *If $U \leq \alpha(x_n, Z)$ then set $X_{n+1} = Z$. Otherwise set $X_{n+1} = X_n$.*
3. *Replace n by $n+1$ and go to step 1.*

We can express the transition kernel Q by

$$\begin{aligned} Q(x, A) &= \int 1(y \in A)\alpha(x,y)g(x,y)\nu(dy) + \left\{1 - \int \alpha(x,y)g(x,y)\nu(dy)\right\} I(x,A) \\ &= \int 1(y \in A)\alpha(x,y)g(x,y)\nu(dy) + r(x)I(x,A), \end{aligned} \tag{3}$$

where I is the identity kernel and $r(x) = 1 - \int \alpha(x,y)g(x,y)\nu(dy)$ is the marginal probability that the chain remains at x. We typically define $Q^{(0)}(x, A) \equiv I(x, A) = 1(x \in A)$. The measure $I(x, \cdot)$ does not have a density with respect to ν. To derive (3), do the following decomposition:

$$\begin{aligned} P(X_{n+1} \in A | X_n = x) =& P(X_{n+1} \in A, U \leq \alpha(x, Z) | X_n = x) \\ &+ P(X_{n+1} \in A, U > \alpha(x, Z) | X_n = x) \\ =& T_1 + T_2. \end{aligned}$$

Then

$$\begin{aligned} T_1 &= E\left[E\left\{1(Z \in A, U \leq \alpha(x,Z)\right) | Z, X_n = x\right\} | X_n = x] \\ &= E\left[\int 1(Z \in A)1(u \leq \alpha(x,Z))du \middle| X_n = x\right] \\ &= E\left[1(Z \in A)\alpha(x,Z) | X_n = x\right] \\ &= \int 1(z \in A)\alpha(x,z)g(x,z)\nu(dz) \end{aligned}$$

and

$$\begin{aligned} T_2 &= E[E\{1(x \in A, U > \alpha(x,Z) | Z, X_n = x\} | X_n = x] \\ &= E[\int 1(x \in A)1(u > \alpha(x,Z))du | X_n = x] \\ &= E[1(x \in A)(1 - \alpha(x,Z)) | X_n = x] \\ &= 1(x \in A) - 1(x \in A)\int \alpha(x,z)g(x,z)\nu(dz). \end{aligned}$$

Given the transition kernel expression in (3), we will show the chain is reversible with respect to P_X, i.e.

$$\iint h(x,y)Q(x,dy)P_X(dx) = \iint h(x,y)Q(y,dx)P_X(dy). \tag{4}$$

for all h such that both sides of (4) exist. Recall that $r(x) = \left\{1 - \int \alpha(x,z)g(x,z)\nu(dz)\right\}$. Starting with the left hand side of (4),

$$
\begin{aligned}
\iint h(x,y)Q(x,dy)P_X(dx) &= \iint h(x,y)\alpha(x,y)g(x,y)\nu(dy)f(x)\nu(dx) \\
&\quad + \iint h(x,y)I(x,dy)r(x)f(x)\nu(dx) \\
&= \iint h(x,y)\min\{g(x,y)f(x), f(y)g(y,x)\}\nu(dy)\nu(dx) \\
&\quad + \int h(x,x)r(x)f(x)\nu(dx) \\
&= \iint h(x,y)\min\{g(y,x)f(y), f(x)g(x,y)\}\nu(dx)\nu(dy) \\
&\quad + \int h(y,y)r(y)f(y)\nu(dy) \\
&= \iint h(x,y)\alpha(y,x)g(y,x)\nu(dx)f(y)\nu(dy) \\
&\quad + \int h(y,y)r(y)f(y)\nu(dy) \\
&= \iint h(x,y)Q(y,dx)P_X(dy).
\end{aligned}
$$

So we have shown that the chain is reversible with respect to P_X, which implies that P_X is the invariant distribution. We also have that if P_X is positive on $\mathcal{X}$ and there exist $\epsilon, R > 0$ such that

$$\inf_{y \in B(x,R)} \min(g(x,y), g(y,x)) > \epsilon,$$

for all $x \in \mathcal{X}$, then Q is P_X-irreducible and aperiodic, so P_X is the unique invariant distribution and we have convergence in the TV norm.

5.2 Random walk Metropolis

If the density of our proposal kernel $g(x,\cdot)$ is of the form $g(x,y) = g^*(y-x)$, where g^* is a density, then we call the resulting MH algorithm a *random walk Metropolis* algorithm. In this case we can write the Hastings ratio in (2) as

$$\frac{f(y)g(y,x)}{f(x)g(x,y)} = \frac{f(y)g^*(x-y)}{f(x)g^*(y-x)}.$$

If g^* is symmetric, i.e. $g^*(-u) = g^*(u)$, then the Hastings ratio simplifies:

$$\frac{f(y)g(y,x)}{f(x)g(x,y)} = \frac{f(y)g^*(y-x)}{f(x)g^*(y-x)} = \frac{f(y)}{f(x)}.$$

e.g. $\mathcal{X} = \mathbb{R}^p$ and $G(x,\cdot) = N_p(x,\Sigma)$; $\mathcal{X} = \mathbb{R}$ and $G(x,\cdot) = \mathrm{Unif}(x-b, x+b)$.

5.2.1 Example: approximately sample from $N(\mu, \sigma^2)$

We give a simple univariate illustration of the random walk Metropolis algorithm to approximately sample from $N(\mu, \sigma^2)$. We will use the symmetric kernel $G(x, \cdot) \sim \text{Unif}(x-b, x+b)$. Then $g(x, y) \propto 1\{y \in (x-b, x+b)\} = 1\{y-x \in (-b, b)\}$ and $f(x) \propto \exp\{-(x-\mu)^2/(2\sigma^2)\}$. Then

$$\alpha(x, y) = \min\left\{1, \frac{f(y)}{f(x)}\right\} = \min\left\{1, \exp\left[\frac{1}{2\sigma^2}\left\{(x-\mu)^2 + (y-\mu)^2\right\}\right]\right\}.$$

We pick $X_0 \in \mathbb{R}$, set $n = 0$, and perform the following steps:

1. Given $X_n = x_n$, generate $Z \sim \text{Unif}(x_n - b, x_n + b)$ and independently generate $U \sim \text{Unif}(0, 1)$.
2. If $U \leq \alpha(x_n, Z)$ then set $X_{n+1} = Z$. Otherwise set $X_{n+1} = X_n$.
3. Replace n by $n + 1$ and go to step 1.

5.2.2 Example: approximately sample from a bivariate posterior

Suppose we measured heights $x_1, \ldots x_n$. Assume that $x = (x_1, \ldots, x_n)'$ is a realization of Z, where

$$\begin{aligned}
(X|M = \mu, V = v) &\sim N_n(\mu 1_n, vI_n)\\
(M|V = v) &\sim N(\mu_M, v_M)\\
V &\sim \text{InvGam}(\alpha, \beta)
\end{aligned}$$

The density for $\text{InvGam}(\alpha, \beta)$ evaluated at v is proportional to $v^{-(\alpha+1)}e^{-\beta/v}$. Our posterior is

$$\begin{aligned}
f(\mu, v|x) &\propto v^{-(\alpha+1)}e^{-\beta/v}v^{-n/2}\exp\left\{-\frac{1}{2v}\sum_{i=1}^{n}(x_i-\mu)^2\right\}\exp\left\{-\frac{1}{2v_M}(\mu-\mu_M)^2\right\}\\
&= v^{-(\alpha+1+n/2)}\exp\left\{-\frac{\beta}{v} - \frac{1}{2v}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{1}{2v_M}(\mu-\mu_M)^2\right\}.
\end{aligned}$$

For illustration, we will use a random walk Metropolis algorithm with a $N_2\{(m_n, v_n)', D\}$ trial distribution to approximately sample from $(M, V|X = x)$. This is an inefficient approach. The covariance matrix will be diagonal: $D = \text{diag}(s_1^2, s_2^2)$. Also,

$$\log \alpha\{(\mu_x, v_x), (\mu_y, v_y)\} = \min\left\{0, \log f(\mu_y, v_y) - \log f(\mu_x, v_x)\right\},$$

where

$$\log f(\mu_y, v_y) = -(\alpha + 1 + n/2)\log v_y - \frac{1}{2v_y}\sum_{i=1}^{n}(x_i - \mu_y)^2 - \frac{1}{2v_M}(\mu_y - \mu_M)^2 - \beta/v_y,$$

provided that $v_y > 0$, otherwise we define $\log f(\mu_y, v_y) = -\infty$. To summarize the algorithm, we pick $(\mu_0, v_0) \in \mathbb{R}^2$ and $D = \text{diag}(s_1^2, s_2^2)$; set $n = 0$; and perform the following steps:

1. Generate $Z \sim N_2\{(\mu_n, v_n)', D\}$ and independently generate $U \sim \mathrm{Unif}(0,1)$.
2. If $\log U \le \log \alpha\{(\mu_n, v_n), Z\}$ then set $(\mu_{n+1}, v_{n+1})' = Z$. Otherwise set $(\mu_{n+1}, v_{n+1})' = (\mu_n, v_n)'$.
3. Replace n by $n+1$ and go to step 1.

5.3 Independence sampler

If the density of our proposal kernel $g(x,\cdot)$ is of the form $g(x,y) = g^*(y)$, where g^* is a density, then we call the resulting MH algorithm an *independence sampler*. In this case we can write the ratio in (2) as

$$\frac{f(y)g(y,x)}{f(x)g(x,y)} = \frac{f(y)g^*(x)}{f(x)g^*(y)}.$$

We need to select g^* carefully to ensure that the Markov chain converges to the target distribution.

5.4 Variable at a time Metropolis–Hastings and the Gibbs sampler

Suppose that there are intermediate $\mathcal{X}$-valued random variables in the Markov chain:

$$\ldots, X_n, X_{n,b}, X_{n+1}, X_{n+1,b}, X_{n+1}, \ldots$$

From X_n to $X_{n,b}$ we use kernel Q_1 and from $X_{n,b}$ to X_{n+1} we use kernel Q_2. The kernel Q for the combination/composition of Q_1 and Q_2 is

$$\begin{aligned}
Q(x_n, A) &= P(X_{n+1} \in A | X_n = x_n) \\
&= E[E\{1(X_{n+1} \in A) | X_{n,b}, X_n = x_n\} | X_n = x_n] \\
&= \int \left\{ \int 1(x_{n+1} \in A) Q_2(x_{n,b}, dx_{n+1}) \right\} Q_1(x_n, dx_{n,b}) \\
&= \int Q_2(x_{n,b}, A) Q_1(x_n, dx_{n,b}) \\
&\equiv Q_1 Q_2(x_n, A)
\end{aligned}$$

Some algorithms for MCMC involve creating a transition kernel Q from sub-transition kernels $Q_1, \ldots, Q_p$. So long has each Q_j has invariant distribution P_X, i.e. $P_X Q_j = P_X$, then $P_X Q_1 \cdots Q_j = P_X Q = P_X$ because $(P_X Q_1) Q_2 = P_X Q_2 = P_X$, so $((P_X Q_1) Q_2) Q_3 = P_X Q_3 = P_X$, and so on.

Suppose that $X_n = (X_{n1}, \ldots, X_{np})$. In *variable at a time Metropolis–Hastings*, we update X_{nj} with the proposal density $g_j(x,\cdot)$ for $j = 1, \ldots, p$. Define

$$\alpha_j(x,y) = \min\left(1, \frac{f(x^{(y,j)}) g_j(x^{(y,j)}, x_j)}{f(x) g_j(x,y)}\right), \tag{5}$$

where $x^{(y,j)} = (x_1, \ldots, x_{j-1}, y, x_{j+1}, \ldots, x_p)$. This denominator is always positive if we start the chain at x with $f(x) > 0$.

Algorithm 2 (Variable at a time Metropolis–Hastings)**.** *Pick or generate* $X_0 \in \mathcal{X}$ *with* $f(X_0) > 0$ *and set* $n = 0$

1. *Set* $Y = (Y_1, \dots, Y_p) = X_n = (X_{n1}, \dots, X_{np})$.
2. *Generate* $Z \sim G_1(Y, \cdot)$ *and independently generate* $U \sim \text{Unif}(0,1)$. *If* $U \le \alpha_1(Y, Z)$ *then set* $Y_1 = Z$ *Otherwise set* $Y_1 = X_{n1}$.
3. *Generate* $Z \sim G_2(Y, \cdot)$ *and independently generate* $U \sim \text{Unif}(0,1)$. *If* $U \le \alpha_2(Y, Z)$ *then set* $Y_2 = Z$ *Otherwise set* $Y_2 = X_{n2}$.
4. $\dots$
5. *Generate* $Z \sim G_p(Y, \cdot)$ *and independently generate* $U \sim \text{Unif}(0,1)$. *If* $U \le \alpha_p(Y, Z)$ *then set* $Y_p = Z$ *Otherwise set* $Y_p = X_{np}$.
6. *Set* $X_{n+1} = Y$*, replace* n *by* $n+1$ *and go to step 1.*

The *Gibbs sampler* is a special case of Algorithm 2. Let x_{-j} be the vector x with its jth entry removed Take $g_j(x, y) = f_j(y|x_{-j})$, where f_j is the density for the conditional distribution of the jth variable in $X \sim f$ given the others. In this case $\alpha_j(x, y) = 1$ because the ratio in (5) is

$$\frac{f(x^{(y,j)})g_j(x^{(y,j)}, x_j)}{f(x)g_j(x, y)} = \frac{f(x^{(y,j)})f_j(x_j|x_{-j})}{f(x)f_j(y|x_{-j})} = \frac{f(x^{(y,j)})K_j f(x)}{f(x)K_j f(x^{(y,j)})} = 1.$$

So the Gibbs sampler always accepts. Let $Y = (Y_1, \dots, Y_p)$ be a random vector with the target distribution (with density f). The Gibbs sampler (with a fixed scan) simplifies to the following:

Algorithm 3 (Gibbs sampler)**.** *Pick or generate* $x_0 = (x_{01}, \dots, x_{0p}) \in \mathcal{X}$ *with* $f(x_0) > 0$ *and set* $n = 0$

1. *Generate a realization* $x_{n+1,1}$ *of* $(Y_1|Y_2 = x_{n2}, \dots, Y_p = x_{np})$.
2. *Generate a realization* $x_{n+1,2}$ *of* $(Y_2|Y_1 = x_{n+1,1}, Y_3 = x_{n,3}, \dots, Y_p = x_{np})$.
3. ...
4. *Generate a realization* $x_{n+1,p}$ *of* $(Y_p|Y_1 = x_{n+1,1}, \dots, Y_{p-1} = x_{n+1,p-1})$.
5. *Replace* n *by* $n+1$ *and go to step 1.*

It is important that the steps in Algorithm 3 are performed sequentially and not in parallel. We can modify the order of this sequence.

5.4.1 Example: approximately sample from a bivariate posterior

Let $x = (x_1, \ldots, x_n)$ be measurements of some numerical characteristic on n subjects. Suppose that x is a realization of X, where

$$\begin{aligned}(X|M = \mu, V = v) &\sim N_n(\mu 1_n, vI_n),\\ (M|V = v) &\sim N(\mu_M, v_M),\\ V &\sim \text{InvGam}(\alpha, \beta),\end{aligned}$$

where $\mu_M, v_M, \alpha, \beta$ are user-specified prior parameters. The density for InvGam(α, β) evaluated at v is proportional to $v^{-(\alpha+1)}e^{-\beta/v}$. The joint density is defined by

$$f(\mu, v, x) \propto v^{-(\alpha+1)}e^{-\beta/v}v^{-n/2} \exp\left\{-\frac{1}{2v}\sum_{i=1}^{n}(x_i - \mu)^2\right\} \exp\left\{-\frac{1}{2v_M}(\mu - \mu_M)^2\right\}.$$

Then

$$\begin{aligned}f(v|\mu, x) &\propto v^{-n/2}e^{-\beta/v} \exp\left\{-\frac{1}{2v}\sum_{i=1}^{n}(x_i - \mu)^2\right\} v^{-(\alpha+1)}\\ &= v^{-(\alpha+1+n/2)} \exp\left[-\left\{\beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right\}/v\right].\end{aligned}$$

Thus

$$(V|M = \mu, X = x) \sim \text{InvGam}\left(\alpha + \frac{n}{2}, \quad \beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right)$$

We also have that

$$\begin{aligned}f(\mu|v, x) &\propto \exp\left\{-\frac{1}{2v}\sum_{i=1}^{n}(x_i - \mu)^2\right\} \exp\left\{-\frac{1}{2v_M}(\mu - \mu_M)^2\right\}\\ &\propto \exp\left(-0.5v^{-1}\sum_{i=1}^{n}x_i^2 + \mu v^{-1}\sum_{i=1}^{n}x_i - 0.5v^{-1}n\mu^2 - 0.5v_M^{-1}\mu^2 + \mu v_M^{-1}\mu_M\right)\\ &\propto \exp\left\{-0.5\mu^2(nv^{-1} + v_M^{-1}) + \mu\left(v^{-1}n\bar{x} + v_M^{-1}\mu_M\right)\right\}\end{aligned}$$

The density for $N(m, w)$ is proportional in y to $\exp\{-0.5y^2v^{-1} + ymw^{-1}\}$, so

$$(M|V = v, X = x) \sim N\left(\frac{n\bar{x} + vv_M^{-1}\mu_M}{n + vv_M^{-1}}, \frac{1}{nv^{-1} + v_M^{-1}}\right)$$

To summarize, we pick (v_0, m_0) and set $n = 0$.

1. Generate v_{n+1} from $(V|M = m_n, X = x)$.
2. Generate m_{n+1} from $(M|V = v_{n+1}, X = x)$.
3. Replace n by $n + 1$ and go to step 1.

5.4.2 Example: fit a linear random effects model

Let y_{it} be the observed response for the ith subject at time t. Suppose that y_{it} is a realization of Y_{it}, where

$$
\begin{aligned}
(Y_{it}|M = m, A = a, V_E = v_E, V_A = v_A) &\sim_{\text{ind}} N(m + a_i, v_E), \\
& \qquad i = 1, \ldots, n, \quad t = 1, \ldots, T_i \\
(M|A = a, V_E = v_e, V_A = v_a) &\sim N(0, v_M) \\
(A|V_E = v_E, V_A = v_A) &\sim N_n(0, v_A I_n) \\
(V_E|V_A = v_A) &\sim \text{InvGam}(\alpha_E, \beta_E) \\
V_A &\sim \text{InvGam}(\alpha_A, \beta_A)
\end{aligned}
$$

Let y be a realization of $Y = (Y_{11}, \ldots, Y_{nT_n})$. Recall that the density for InvGam(α, β) evaluated at v is proportional to $v^{-(\alpha+1)}e^{-\beta/v}$. The joint density is defined by

$$
\begin{aligned}
f(&y,a, m, v_E, v_A) \\
&\propto v_A^{-(\alpha_A+1)} e^{-\beta_A/v_A} v_E^{-(\alpha_E+1)} e^{-\beta_E/v_E} \\
&\quad * v_A^{-n/2} \exp\left(-\frac{1}{2v_A}\sum_{i=1}^{n} a_i^2\right) v_M^{-1/2} \exp\left(-\frac{1}{2v_M} m^2\right) \\
&\quad * v_E^{-\sum_{i=1}^{n} T_i/2} \exp\left\{-\frac{1}{2v_E}\sum_{i=1}^{n}\sum_{t=1}^{T_i}(y_{it} - m - a_i)^2\right\}.
\end{aligned}
$$

So we have that

$$
\begin{aligned}
f(v_A|a, m, v_E, y) &\propto v_A^{-(\alpha_A+1+n/2)} \exp\left\{-\frac{1}{v_A}\left(\beta_A + \frac{1}{2}\sum_{i=1}^{n} a_i^2\right)\right\} \\
f(v_E|a, m, v_A, y) &\propto v_E^{-\left(\alpha_E+1+\frac{1}{2}\sum_{i=1}^{n} T_i\right)} \exp\left[-\frac{1}{v_E}\left\{\beta_E + \frac{1}{2}\sum_{i=1}^{n}\sum_{t=1}^{T_i}(y_{it} - m - a_i)^2\right\}\right]
\end{aligned}
$$

Thus

$$
\begin{aligned}
(V_A|A = a, M = m, V_E = v_E, Y = y) &\sim \text{InvGam}\left(\alpha_A + n/2, \quad \beta_A + \frac{1}{2}\sum_{i=1}^{n} a_i^2\right) \\
(V_E|A = a, M = m, V_A = v_A, Y = y) &\sim \text{InvGam}\left(\alpha_E + \frac{1}{2}\sum_{i=1}^{n} T_i, \quad \beta_E + \frac{1}{2}\sum_{i=1}^{n}\sum_{t=1}^{T_i}(y_{it} - m - a_i)^2\right)
\end{aligned}
$$

We also have that

$$
\begin{aligned}
f(m|a, v_A, v_E, y) &\propto \exp\left\{-\frac{1}{2v_E}\sum_{i=1}^{n}\sum_{t=1}^{T_i}(y_{it}-a_i-m)^2\right\}\exp\left(-\frac{1}{2v_M}m^2\right) \\
&\propto \exp\left[-\frac{1}{2v_E}\sum_{i=1}^{n}\sum_{t=1}^{T_i}\left\{(y_{it}-a_i)^2-2(y_{it}-a_i)m+m^2\right\}-\frac{1}{2v_M}m^2\right] \\
&\propto \exp\left\{-0.5m^2\left(\sum_{i=1}^{n}T_i/v_E+1/v_M\right)+m\sum_{i=1}^{n}\sum_{t=1}^{T_i}(y_{it}-a_i)/v_E\right\}
\end{aligned}
$$

This implies that

$$
(M|A=a, V_E=v_E, V_A=v_A, Y=y) \sim N\left(\frac{\sum_{i=1}^{n}\sum_{t=1}^{T_i}(y_{it}-a_i)}{\sum_{i=1}^{n}T_i+v_E/v_M}, \frac{1}{\sum_{i=1}^{n}T_i/v_E+1/v_M}\right)
$$

Let a_{-i} be the vector a with its ith entry removed. We also have that

$$
\begin{aligned}
f(a_i|a_{-i}, m, v_A, v_E, y) &\propto \exp\left\{-\frac{1}{2v_E}\sum_{t=1}^{T_i}(y_{it}-m-a_i)^2\right\}\exp\left(-\frac{1}{2v_A}a_i^2\right) \\
&\propto \exp\left[-\frac{1}{2v_E}\sum_{t=1}^{T_i}\{(y_{it}-m)^2-2a_i(y_{it}-m)+a_i^2\}-\frac{1}{2v_A}a_i^2\right] \\
&\propto \exp\left\{-0.5a_i^2(T_i/v_E+1/v_A)+a_i\sum_{t=1}^{T_i}(y_{it}-m)/v_E\right\}
\end{aligned}
$$

This implies that

$$
(A_i|M=m, A_{-i}=a_{-i}, V_E=v_E, V_A=v_A, Y=y) \sim N\left(\frac{\sum_{t=1}^{T_i}(y_{it}-m)}{T_i+v_E/v_A}, \frac{1}{T_i/v_E+1/v_A}\right),
$$

for $i = 1, \ldots, n$.

To summarize, we pick $(m_0, a_0, v_{A,0}, v_{E,0})$ and set $n = 0$.

1. Generate m_{n+1} from $(M|A=a_n, V_E=v_{E,n}, V_A=v_{A,n}, Y=y)$.
2. Generate $a_{n+1,i}$ from $(A_i|M=m_{n+1}, V_E=v_{E,n}, V_A=v_{A,n}, Y=y)$ for $i=1,\ldots,n$ in parallel.
3. Generate $v_{A,n+1}$ from $(V_A|A=a_{n+1}, M=m_{n+1}, V_E=v_{E,n}, Y=y)$.
4. Generate $v_{E,n+1}$ from $(V_E|A=a_{n+1}, M=m_{n+1}, V_A=v_{A,n+1}, Y=y)$
5. Replace n by $n+1$ and go to step 1.

5.4.3 Example: Bayesian ridge regression

Let y_i be the measured response for the ith case and let $x_i = (1, x_{i2}, \ldots, x_{ip})' \in \mathbb{R}^p$ be the values of the predictors for the ith case. Define $X \in \mathbb{R}^{n\times p}$ to have ith row x_i. Assume that $(y_1, \ldots, y_n)'$ is a realization of Y, where

$$\begin{aligned}
(Y|B=\beta, V=v, L=\lambda) &\sim N_n(X\beta, vI_n)\\
(B|V=v, L=\lambda) &\sim N_p\left(\tilde{\beta}, \frac{v}{\lambda}I_p\right)\\
(V|L=\lambda) &\sim \text{InvGam}(a_V, b_V)\\
L &\sim \text{Gamma}(a_L, b_L)
\end{aligned}$$

The Gamma(a,b) distribution has density at λ proportional to $\lambda^{a-1}e^{-b\lambda}$. Then

$$\begin{aligned}
f(\beta, v, \lambda|y) &\propto v^{-a_V+1}e^{-b_V/v}\lambda^{a_L-1}e^{-b_L\lambda}\lambda^{p/2}v^{-p/2}\exp\left\{-\frac{\lambda}{2v}(\beta-\tilde{\beta})'(\beta-\tilde{\beta})\right\}\\
&\quad * v^{-n/2}\exp\left\{-\frac{1}{2v}(y-X\beta)'(y-X\beta)\right\}\\
&= \lambda^{a_L-1+p/2}v^{-(a_V+1+p/2+n/2)}\\
&\quad * \exp\left\{-\frac{1}{2v}(y-X\beta)'(y-X\beta)-\frac{\lambda}{2v}(\beta-\tilde{\beta})'(\beta-\tilde{\beta})-b_L\lambda-\frac{b_V}{v}\right\}
\end{aligned}$$

We will use the fact that $\phi(x, \mu, \Sigma) \propto_x \exp(-0.5x'\Sigma^{-1}x + x'\Sigma^{-1}\mu)$. We have that

$$\begin{aligned}
f(\beta|v, \lambda, y) &\propto \exp\left\{-\frac{1}{2v}(y-X\beta)'(y-X\beta)-\frac{\lambda}{2v}(\beta-\tilde{\beta})'(\beta-\tilde{\beta})-b_L\lambda-\frac{b_V}{v}\right\}\\
&\propto \exp\left\{-\frac{1}{2}\beta'\left(\frac{1}{v}X'X+\frac{\lambda}{v}I_p\right)\beta+\beta'\left(\frac{1}{v}X'y+\frac{\lambda}{v}\tilde{\beta}\right)\right\},
\end{aligned}$$

which implies that

$$(B|V=v, L=\lambda, Y=y) \sim N\left((X'X+\lambda I_p)^{-1}(X'y+\lambda\tilde{\beta}), \quad v(X'X+\lambda I_p)^{-1}\right).$$

Let $\mu_B = (X'X+\lambda I_p)^{-1}(X'y+\lambda\tilde{\beta})$. Then

$$\begin{aligned}
f(\beta|v, \lambda, y) &\propto_{\beta,v} v^{-p/2}\exp\left\{-\frac{1}{2v}(\beta-\mu_B)'(X'X+\lambda I_p)(\beta-\mu_B)\right\}\\
&\propto_{\beta,v} v^{-p/2}\exp\left[-\frac{1}{2v}\{\beta'(X'X+\lambda I_p)\beta-2\beta'(X'X+\lambda I_p)\mu_B+\mu_B'(X'X+\lambda I_p)\mu_B\}\right]\\
&\propto_{\beta,v} v^{-p/2}\exp\left[-\frac{1}{2v}\left\{\beta'(X'X+\lambda I_p)\beta-2\beta'(X'y+\lambda\tilde{\beta})+\mu_B'(X'X+\lambda I_p)\mu_B\right\}\right]
\end{aligned}$$

We can write

$$
\begin{aligned}
f(\beta, v|\lambda, y) \propto\ & v^{-(a_V+1+n/2)} v^{-p/2} \\
& * \exp\left[-\frac{1}{2v}\left\{y'y - 2\beta'X'y + \beta'X'X\beta + \lambda\beta'\beta - 2\lambda\beta'\tilde{\beta} + \lambda\tilde{\beta}'\tilde{\beta} + 2b_V\right\}\right] \\
=& f(\beta|v, \lambda, y) v^{-(a_V+1+n/2)} \exp\left[-\frac{1}{2v}\left\{y'y - \mu_B'(X'X + \lambda I_p)\mu_B + \lambda\tilde{\beta}'\tilde{\beta} + 2b_V\right\}\right] \\
\propto& f(\beta|v, \lambda, y) f(v|\lambda, y)
\end{aligned}
$$

which implies that

$$
(V|L = \lambda, Y = y) \sim \text{InvGam}\left(a_V + \frac{n}{2}, \quad b_V + \frac{1}{2}\left\{y'y - \mu_B'(X'X + \lambda I_p)\mu_B + \lambda\tilde{\beta}'\tilde{\beta}\right\}\right)
$$

We have that

$$
f(\lambda|\beta, v, y) \propto \lambda^{a_L - 1 + p/2} \exp\left\{-\lambda\left(\frac{1}{2v}\|\beta - \tilde{\beta}\|^2 + b_L\right)\right\},
$$

which implies that

$$
(L|B = \beta, V = v, Y = y) \sim \text{Gamma}\left(a_L + \frac{p}{2}, \quad \frac{1}{2v}\|\beta - \tilde{\beta}\|^2 + b_L\right).
$$

To summarize, we pick $(\beta_0, v_0, \lambda_0)$ and set $n = 0$.

1. Generate (β_{n+1}, v_{n+1}) from $(B, V|L = \lambda_n, Y = y)$ with the following two steps:
   (a) Generate v_{n+1} from $(V|L = \lambda_n, Y = y)$.
   (b) Generate β_{n+1} from $(B|V = v_{n+1}, L = \lambda_n, Y = y)$
2. Generate λ_{n+1} from $(L|B = \beta_{n+1}, V = v_{n+1}, Y = y)$.
3. Replace n by $n + 1$ and go to step 1.

6 Inference for $E\{h(X)\}$, where X has the target distribution

This section is based on Atchade (2008) and Jones (2013). Suppose our goal is to estimate $E\{h(X)\}$ where $X \sim P_X$ (the target distribution with density f). Using the Markov chain $X_0, X_1, \ldots$ with invariant distribution P_X (from our sampling algorithm), we estimate $E\{h(X)\}$ with

$$
\widehat{E}\{h(X)\} = \frac{1}{n}\sum_{i=1}^{n} h(X_i).
$$

If the chain is P_X-irreducible and aperiodic, then $\widehat{E}\{h(X)\} \to E\{h(X)\}$ almost surely as $n \to \infty$. A Markov chain with transition kernel Q is *geometrically ergodic* if there exists a $\rho \in (0, 1)$ and a function $M : \mathcal{X} \to [0, \infty)$ such that

$$\|Q^{(n)}(x, \cdot) - P_X\|_{\text{TV}} \leq M(x)\rho^n,$$

for all $x \in \mathcal{X}$ and all non-negative integers n. Here are two selected results (Jones et al., 2006; Jones, 2013):

- If the Harris chain is geometrically ergodic and $E\{|h(X)|^{2+\delta}\} < \infty$ for some $\delta > 0$, then
$$\sqrt{n}\left[\widehat{E}\{h(X)\} - E\{h(X)\}\right] \xrightarrow{L} N(0, v_h).$$
- If the Harris chain is geometrically ergodic, reversible, and $E\{h^2(X)\} < \infty$, then
$$\sqrt{n}\left[\widehat{E}\{h(X)\} - E\{h(X)\}\right] \xrightarrow{L} N(0, v_h).$$

In general, $\text{var}\left[\widehat{E}\{h(X)\}\right] = \sum_{i=1}^n \sum_{j=1}^n \text{cov}\{h(X_i), h(X_j)\}$. Without giving details on how to simplify this formula, we will estimate the asymptotic variance v_h using the Batch means method. Suppose we have b batches of size k. Let

$$Y_j = \frac{1}{k} \sum_{i=(j-1)k+1}^{jk} h(X_i), \quad j = 1, \ldots, b.$$

The batch means estimator of v_h is

$$\hat{v}_h = \frac{k}{b-1} \sum_{j=1}^{b} \left[Y_j - \widehat{E}\{h(X)\}\right]^2.$$

This estimator is inconsistent in general, but is consistent if we let k and b increase with n and assume that other regularity conditions hold (Jones, 2013). Jones suggested that $k = \sqrt{n}$ and $b = n/k$ are sensible choices. Assuming that $\widehat{E}\{h(X)\}$ is approximately Normal, a $100(1-\alpha)\%$ approximate confidence interval for $E\{h(X)\}$ is

$$\widehat{E}\{h(X)\} \pm \texttt{qnorm}(1 - \alpha/2)\sqrt{\frac{\hat{v}_h}{n}}.$$

References

Atchade, Y. F. (2008). Course notes for Statistics 606. Lecture notes.

Jones, G. L. (2013). Course notes for STAT 8701: Computational statistical methods. Lecture notes.

Jones, G. L., Haran, M., Caffo, B. S., and Neath, R. (2006). Fixed-width output analysis for markov chain monte carlo. *Journal of the American Statistical Association*, 101:1537–1547.

Keener, R. W. (2005). Statistical theory: A medley of core topics. Notes for a course in theoretical statistics.