7.1 Introduction

Chapter 6 described some principles for simulation based on Markov chains, as well as some implementation directions, including the generic random walk Metropolis–Hastings algorithm. This chapter extends the scope of MCMC algorithms by studying another class of now-common MCMC methods, called Gibbs sampling. The appeal of those specific algorithms is that first they gather most of their calibration from the target density and second they allow us to break complex problems (such as high dimensional target distributions, for which a random walk Metropolis–Hastings algorithm is almost impossible to build) into a series of easier problems, like a sequence of small-dimension targets. There may be caveats to this simplification in that the sequence of simple problems may take *in fine* a long time to converge, but Gibbs sampling is nonetheless an interesting candidate when dealing with a new problem.

The name *Gibbs sampling* comes from the landmark paper by Geman and Geman (1984), which first applied a Gibbs sampler on a *Gibbs random field*. For good or bad, it then stuck despite this weak link. Indeed, it is in fact a special case of the Metropolis–Hastings algorithm as detailed in Robert and Casella (2004, Section 10.6.1). The work of Geman and Geman (1984), built on that of Metropolis et al. (1953), Hastings (1970) and Peskun (1973), influenced Gelfand and Smith (1990) to write a paper that sparked new interest in Bayesian methods, statistical computing, algorithms, and stochastic processes through the use of computing algorithms such as the Gibbs sampler and the Metropolis–Hastings algorithm. It is interesting to see, in retrospect, that earlier papers such as Tanner and Wong (1987) and Besag and Clifford (1989) had proposed similar solutions (but did not receive the same response from the statistical community).

7.2 The two-stage Gibbs sampler

The *two-stage Gibbs sampler* creates a Markov chain from a joint distribution in the following way. If two random variables X and Y have joint density $f(x, y)$, with corresponding conditional densities $f_{Y|X}$ and $f_{X|Y}$, the two-stage Gibbs sampler generates a Markov chain (X_t, Y_t) according to the following steps:

Algorithm 7 Two-stage Gibbs sampler
`Take` $X_0 = x_0$
`For` $t = 1, 2, \ldots,$ `generate`
1. $Y_t \sim f_{Y|X}(\cdot|x_{t-1})$;
2. $X_t \sim f_{X|Y}(\cdot|y_t)$.

Algorithm 7 is then straightforward to implement as long as simulating from both conditionals is feasible.[1] It is also easy to see why, if (X_t, Y_t) is distributed from f, then so is (X_{t+1}, Y_{t+1}), because both steps of iteration t use simulation from the true conditionals. Convergence of the Markov chain (and thus the algorithm) is therefore ensured unless the supports of the conditionals are not connected.

Example 7.1. To start with an obvious illustration, consider the bivariate normal model

$$(X, Y) \sim \mathcal{N}_2\left(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right), \tag{7.1}$$

for which the Gibbs sampler is
`Given` x_t`, generate`

$$Y_{t+1} \mid x_t \sim \mathcal{N}(\rho x_t,\ 1-\rho^2),$$
$$X_{t+1} \mid y_{t+1} \sim \mathcal{N}(\rho y_{t+1},\ 1-\rho^2).$$

The subchain $(X_t)_t$ then satisfies

$$X_{t+1}|X_t = x_t \sim \mathcal{N}(\rho^2 x_t,\ 1-\rho^4),$$

and a recursion shows that

$$X_t|X_0 = x_0 \sim \mathcal{N}(\rho^{2t} x_0,\ 1-\rho^{4t}),$$

which does indeed converge to $\mathcal{N}(0,1)$ as t goes to infinity. ◀

As illustrated by the example above, the sequence (X_t, Y_t), $t = 1, \ldots, T$, produced by a Gibbs sampler converges to the *joint* distribution f and, as a consequence, both sequences $(X_t)_t$ and $(Y_t)_t$ converge to their respective *marginal distributions*.

Exercise 7.1 Show that the subsequence (X_t) resulting from Algorithm 7 is a Markov chain. (*Hint:* Use the fact that (X_t, Y_t) is generated conditional on X_{t-1} only.)

Perhaps the main reason why the Gibbs sampler became so popular in the 1990s as the reference MCMC algorithm is that it was the perfect computational complement to hierarchical models, which were then starting to be seriously investigated. As detailed and justified in Section 7.5, a hierarchical model specifies a joint distribution as successive layers of conditional distributions. The following example gives a first look at hierarchical models.

[1] When $f(x, y)$ is available in closed form, up to a normalizing constant, so are $f_{Y|X}$ and $f_{X|Y}$. Therefore, if simulating directly from those conditionals is not possible, Monte Carlo or MCMC approximations can be used, as developed in Section 7.6.3.

Example 7.2. Considering the pair of distributions

$$X|\theta \sim \mathcal{B}\text{in}(n,\theta)\,, \quad \theta \sim \mathcal{B}e(a,b),$$

leads to the joint distribution

$$f(x,\theta) = \binom{n}{x} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\, \theta^{x+a-1}(1-\theta)^{n-x+b-1}.$$

The corresponding conditional distribution of $X|\theta$ is given above, while $\theta|x \sim \mathcal{B}e(x+a, n-x+b)$. The associated Gibbs sampler can be implemented as

```
> Nsim=5000                          #initial values
> n=15
> a=3
> b=7
> X=T=array(0,dim=c(Nsim,1))         #init arrays
> T[1]=rbeta(1,a,b)                  #init chains
> X[1]=rbinom(1,n,T[1])
> for (i in 2:Nsim){                 #sampling loop
+   X[i]=rbinom(1,n,T[i-1])
+   T[i]=rbeta(1,a+X[i],n-X[i]+b)
+   }
```

and its output is illustrated in Figure 7.1 for each marginal. Since this is a toy example, the closed-form marginals are available and thus produced on top of the histograms, and they show a good fit for both Gibbs samples. ◀

Exercise 7.2 The marginal distribution of θ in Example 7.2 is the standard $\mathcal{B}e(a,b)$ distribution, but the marginal distribution of X is less standard and is known as the *beta-binomial* distribution.

a. Produce a closed-form expression for the beta-binomial density by integrating $f(x,\theta)$ in Example 7.2 with respect to θ.
b. Use this expression to create the function `betabi` in `R`. Then use the `R` command `curve(betabi(x,a,b,n))` to draw a curve on top of the histogram as in Figure 7.1.

Example 7.3. Consider the posterior distribution on (θ, σ^2) associated with the joint model

$$\begin{aligned} X_i &\sim \mathcal{N}(\theta, \sigma^2), \quad i = 1, \ldots, n, \\ \theta &\sim \mathcal{N}(\theta_0, \tau^2)\,, \quad \sigma^2 \sim \mathcal{IG}(a,b), \end{aligned} \tag{7.2}$$

where $\mathcal{IG}(a,b)$ is the inverted gamma distribution (that is, the distribution of the inverse of a gamma variable), with density $b^a(1/x)^{a+1}e^{-b/x}/\Gamma(a)$ and with

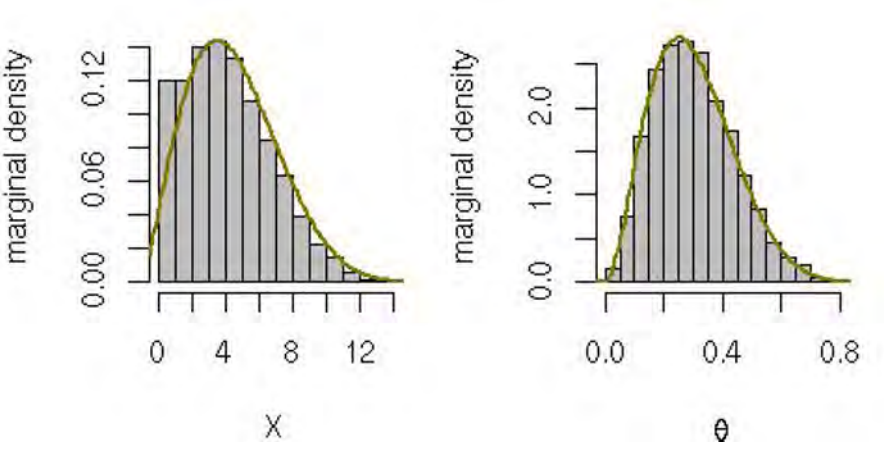

Fig. 7.1. Histograms of marginal distributions from the Gibbs sampler of Example 7.2 based on 5000 iterations of Algorithm 7 for $n = 15, a = 3, b = 7$. The true marginal distribution of θ is $\mathcal{B}e(a, b)$ and the marginal distribution of X is beta-binomial.

θ_0, τ^2, a, b specified. Writing $\mathbf{x} = (x_1, \ldots, x_n)$, the posterior distribution on (θ, σ^2) is given by

$$f(\theta, \sigma^2|\mathbf{x}) \propto \left[\frac{1}{(\sigma^2)^{n/2}} e^{-\sum_i (x_i-\theta)^2/(2\sigma^2)}\right] \times \left[\frac{1}{\tau} e^{-(\theta-\theta_0)^2/(2\tau^2)}\right] \times \left[\frac{1}{(\sigma^2)^{a+1}} e^{1/b\sigma^2}\right], \tag{7.3}$$

from which we can get the full conditionals of θ and σ^2. (Note that this is not a regular conjugate setting in that integrating θ or σ^2 in this density does not produce a standard density.) Writing $\mathbf{x} = (x_1, \ldots, x_n)$, we have

$$\begin{aligned} \pi(\theta|\mathbf{x}, \sigma^2) &\propto e^{-\sum_i (x_i-\theta)^2/(2\sigma^2)} e^{-(\theta-\theta_0)^2/(2\tau^2\sigma^2)}, \\ \pi(\sigma^2|\mathbf{x}, \theta) &\propto \left(\frac{1}{\sigma^2}\right)^{(n+2a+3)/2} e^{-\frac{1}{2\sigma^2}\left(\sum_i (x_i-\theta)^2 + (\theta-\theta_0)^2/\tau^2 + 2/b\right)}. \end{aligned} \tag{7.4}$$

These densities correspond to

$$\theta|\mathbf{x}, \sigma^2 \sim \mathcal{N}\left(\frac{\sigma^2}{\sigma^2 + n\tau^2}\, \theta_0 + \frac{n\tau^2}{\sigma^2 + n\tau^2}\, \bar{x},\ \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}\right)$$

and

$$\sigma^2|\mathbf{x}, \theta \sim \mathcal{IG}\left(\frac{n}{2} + a, \frac{1}{2}\sum_i (x_i - \theta)^2 + b\right),$$

where $\bar{x}$ is the empirical average of the observations, as the full conditional distributions to be used in a Gibbs sampler.

A study on metabolism in 15-year-old females yielded the following data, denoted by $\mathbf{x}$,

```
> x=c(91,504,557,609,693,727,764,803,857,929,970,1043,
+     1089,1195,1384,1713)
```

corresponding to their energy intake, measured in megajoules, over a 24 hour period (also available in the dataset `Energy`). Using the normal model above, with θ corresponding to the true mean energy intake, the Gibbs sampler can be implemented as

```
> xbar=mean(x)
> sh1=(n/2)+a
> sigma=theta=rep(0,Nsim)                    #init arrays
> sigma{1}=1/rgamma(1,shape=a,rate=b)        #init chains
> B=sigma2{1}/(sigma2{1}+n*tau2)
> theta{1}=rnorm(1,m=B*theta0+(1-B)*xbar,sd=sqrt(tau2*B))
> for (i in 2:Nsim){
+   B=sigma2[i-1]/(sigma2[i-1]+n*tau2)
+   theta[i]=rnorm(1,m=B*theta0+(1-B)*xbar,sd=sqrt(tau2*B))
+   ra1=(1/2)*(sum((x-theta[i])^2))+b
+   sigma2[i]=1/rgamma(1,shape=sh1,rate=ra1)
+   }
```

where `theta0`, `tau2`, `a`, and `b` are specified values. The posterior means of θ and σ^2 are 872.402 and $136,229.2$, giving as an estimate of σ 369.092. Histograms of the posterior distributions of $\log(\theta)$ and $\log(\sigma)$ are given in Figure 7.2. ◀

Exercise 7.3 In connection with Example 7.3

a. Reproduce Figure 7.2 and superimpose the true marginal posteriors of $\log(\theta)$ and $\log(\sigma)$ by integrating $f(\theta, \sigma^2|\mathbf{x})$ in σ^2 and θ, respectively.
b. Investigate the sensitivity of the answer for a range of specifications of the hyperparameter values `theta0`, `tau2`, `a`, and `b`. Specifically, compute point estimates and confidence limits for θ and σ over a range of values for those parameters.

We want to point out that recognizing the full conditionals from a joint distribution is not that difficult. For example, the posterior distribution proportional to (7.3) is obtained by multiplying the densities in the specification (7.2).

To find a *full* conditional (that is, the conditional distribution of one parameter conditional on all others), we merely need to pick out all of the terms

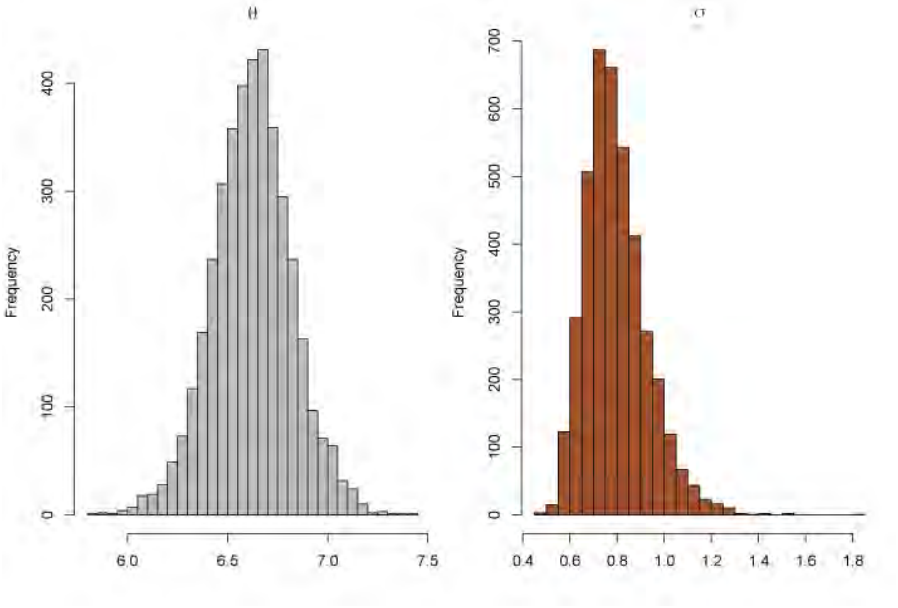

Fig. 7.2. Histograms of marginal posterior distributions of the log-mean and log-standard deviation from the Gibbs sampler of Example 7.3 based on 5000 iterations, with $a = b = 3$, $\tau_2 = 10$ and $\theta_0 = 5$. The 90% interval for $\log(\theta)$ is $(6.299, 6.960)$ and for $\log(\sigma)$ it is $(0.614, 1.029)$.

in the joint distribution that involve that parameter. For example, from (7.3), we see that

$$f(\theta|\sigma^2, \mathbf{x}) \propto \left[\frac{1}{(\sigma^2)^{n/2}} e^{-\sum_i (x_i-\theta)^2/(2\sigma^2)}\right] \times \left[\frac{1}{\tau} e^{-(\theta-\theta_0)^2/(2\tau^2)}\right],$$
$$f(\sigma^2|\theta, \mathbf{x}) \propto \left[\frac{1}{(\sigma^2)^{n/2}} e^{-\sum_i (x_i-\theta)^2/(2\sigma^2)}\right] \times \left[\frac{1}{(\sigma^2)^{a+1}} e^{1/b\sigma^2}\right].$$

It should then be easy to see that the full conditional of σ^2 will be an inverted gamma distribution, as defined on page 202 (see also Exercise 7.19). For θ, although there is a little more algebra involved in the derivation, we can recognize that the full conditional will be normal. See Exercise 7.20 for an illustration with a larger hierarchy.

Exercise 7.4 Make explicit the derivations that connect the expressions above and the full conditional distributions in (7.4).