# 7

# Gibbs Samplers

*"Come, Watson , come!" he cried. "The game is afoot."*

**Arthur Conan Doyle**
*The Adventure of the Abbey Grange*

## Reader's guide

This chapter covers both the two-stage and the multistage Gibbs samplers. Although the former is a special case of the latter, the two-stage sampler has superior convergence properties and applies naturally in a wide range of statistical models that do not call for the generality of the multistage sampler. Nevertheless, the multistage Gibbs sampler enjoys many optimality properties and still might be considered the workhorse of the MCMC world. Following the introduction in Section 7.1 with some background, we develop the two-stage Gibbs sampler in Section 7.2, moving to the multistage Gibbs sampler in Section 7.3. The Gibbs sampler is particularly well-suited to handle experiments with missing data and models with latent variables, as shown in Section 7.4. Although we make use of hierarchical models throughout the chapter, we focus on their processing in Section 7.5. Section 7.6 looks at a number of additional topics such as Rao–Blackwellization, reparameterization, and the effect of using improper priors.

C.P. Robert, G. Casella, *Introducing Monte Carlo Methods with R*, Use R,
DOI 10.1007/978-1-4419-1576-4_7, © Springer Science+Business Media, LLC 2010

## 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 $\theta$ 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 $\theta$.
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 $(\theta, \sigma^2)$ 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

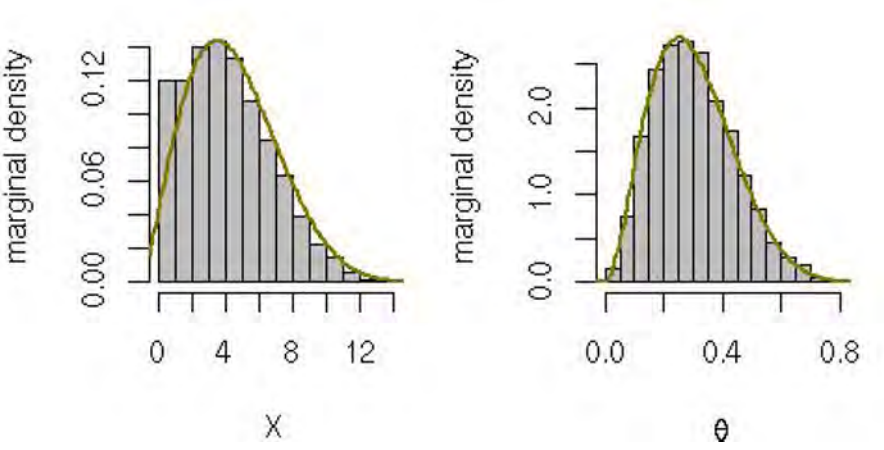


**Fig. 7.1.** Histograms of marginal distributions from the Gibbs sampler of Example 7.2 based on 5000 iterations of Algorithm 7 for $n = 15, a = 3, b = 7$. The true marginal distribution of $\theta$ is $\mathcal{B}e(a, b)$ and the marginal distribution of $X$ is beta-binomial.

$\theta_0, \tau^2, a, b$ specified. Writing $\mathbf{x} = (x_1, \ldots, x_n)$, the posterior distribution on $(\theta, \sigma^2)$ 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 $\theta$ and $\sigma^2$. (Note that this is not a regular conjugate setting in that integrating $\theta$ or $\sigma^2$ 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 $\theta$ 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 $\theta$ and $\sigma^2$ are $872.402$ and $136,229.2$, giving as an estimate of $\sigma$ $369.092$. 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 $\sigma^2$ and $\theta$, 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 $\theta$ and $\sigma$ 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

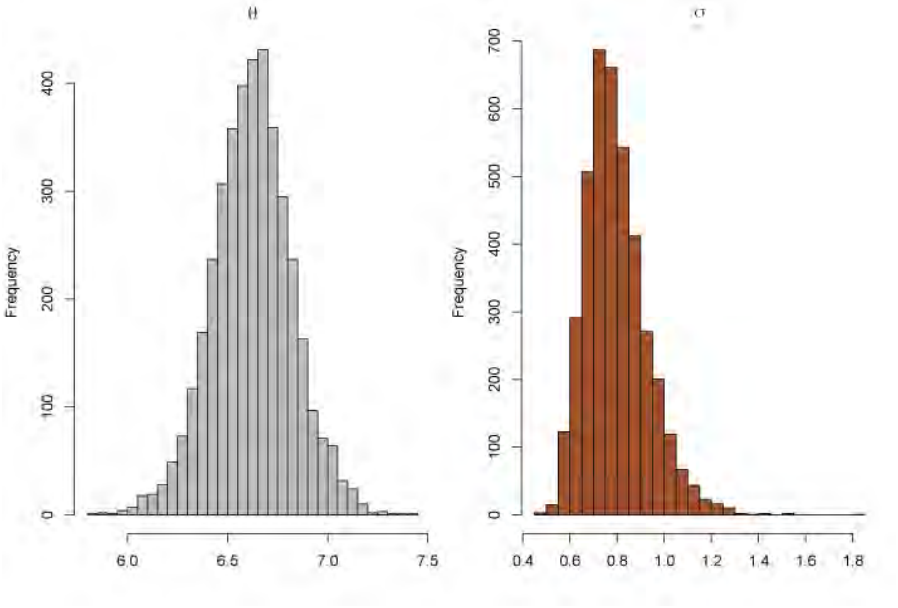


**Fig. 7.2.** Histograms of marginal posterior distributions of the log-mean and log-standard deviation from the Gibbs sampler of Example 7.3 based on 5000 iterations, with $a = b = 3$, $\tau_2 = 10$ and $\theta_0 = 5$. The 90% interval for $\log(\theta)$ is $(6.299, 6.960)$ and for $\log(\sigma)$ it 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 $\sigma^2$ will be an inverted gamma distribution, as defined on page 202 (see also Exercise 7.19). For $\theta$, 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).

## 7.3 The multistage Gibbs sampler

There is a natural extension from the two-stage Gibbs sampler to the general multistage Gibbs sampler. Suppose that, for some $p > 1$, the random variable $\mathbf{X} \in \mathcal{X}$ can be written as $\mathbf{X} = (X_1, \ldots, X_p)$, where the $X_i$'s are either unidimensional or multidimensional components. Moreover, suppose that we can simulate from the corresponding conditional densities $f_1, \ldots, f_p$, that is, we can simulate

$$X_i | x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p \sim f_i(x_i | x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p)$$

for $i = 1, 2, \ldots, p$. The associated *Gibbs sampling* algorithm (or *Gibbs sampler*) is given by the following transition from $X^{(t)}$ to $X^{(t+1)}$:

**Algorithm 8 The Multistage Gibbs Sampler**

At iteration $t = 1, 2, \ldots,$ given $\mathbf{x}^{(t)} = (x_1^{(t)}, \ldots, x_p^{(t)})$, generate

1. $X_1^{(t+1)} \sim f_1(x_1 | x_2^{(t)}, \ldots, x_p^{(t)})$;
2. $X_2^{(t+1)} \sim f_2(x_2 | x_1^{(t+1)}, x_3^{(t)}, \ldots, x_p^{(t)})$;

$\vdots$

p. $X_p^{(t+1)} \sim f_p(x_p | x_1^{(t+1)}, \ldots, x_{p-1}^{(t+1)})$.

The densities $f_1, \ldots, f_p$ are called the *full conditionals*, and a particular feature of the Gibbs sampler is that these are the only densities used for simulation. Thus, even in a high-dimensional problem, *all of the simulations may be univariate*, which is usually an advantage.

**Example 7.4.** As an extension of Example 7.1, consider the multivariate normal density

$$(X_1, X_2, \ldots, X_p) \sim \mathcal{N}_p\left(0, (1-\rho)I + \rho J\right), \tag{7.5}$$

where $I$ is the $p \times p$ identity matrix and $J$ is a $p \times p$ matrix of ones. This is a model for *equicorrelation*, as $\text{corr}(X_i, X_j) = \rho$ for every $i$ and $j$. Using standard formulas for the conditional distributions of a multivariate normal random variable (see, for example, Johnson and Wichern, 1988), it is straightforward but tedious to verify that

$$X_i | x_{(-i)} \sim \mathcal{N}\left(\frac{(p-1)\rho}{1+(p-2)\rho}\bar{x}_{(-i)}, \frac{1+(p-2)\rho-(p-1)\rho^2}{1+(p-2)\rho}\right),$$

where $x_{(-i)} = (x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p)$ and $\bar{x}_{(-i)}$ is the mean of this vector. The Gibbs sampler that generates from these univariate normals can then be easily derived, although it is useless for this problem (Exercise 7.5). It is,

however, a short step to consider the setup where the components of the normal vector are restricted to a subset of $\mathbb{R}^p$. If this subset is a hypercube,

$$\mathfrak{H} = \prod_{i=1}(a_i, b_i)\,,$$

then the corresponding conditionals simply are the normals above restricted to $(a_i, b_i)$ for $i = 1, \ldots, p$ (in which case an exact algorithm such as `sadmvn` can be used). For more complex constraints, a Gibbs sampler is however (almost) required, as exact solutions do not exist. This Gibbs sampler is still based on normal full conditionals, which are now restricted to subsets of the real line and thus easily simulated (Exercise 2.22). ◀

**Exercise 7.5** Given the normal target $\mathcal{N}_p\left(0, (1-\rho)I + \rho J\right)$:

a. Write a Gibbs sampler using the conditional distributions provided in Example 7.4. Run your `R` code for $p = 5$ and $\rho = .25$, and verify graphically that the marginals are all $\mathcal{N}(0, 1)$.
b. Compare your algorithm using $T = 500$ iterations with `rmnorm` described in Section 2.2.1 in terms of execution time.
c. Propose a constrained subset that is not a hypercube, and derive the corresponding Gibbs sampler. (*Hint*: Consider, for example, a constraint such as $\sum_{i=1}^{m} x_i^2 \leq \sum_{i=m+1}^{p} x_i^2$ for $m \leq p-1$.)

Models more complex than the one in Example 7.3 can be considered for the normal sampling model, as in the following case.

**Example 7.5.** A hierarchical specification for the normal model is the *one-way random effects model*. There are different ways to parameterize this model, but a possibility is as follows (see others in Example 7.14 and Exercise 7.24):

$$\begin{aligned} X_{ij} &\sim \mathcal{N}(\theta_i, \sigma^2), \quad i = 1, \ldots, k, \quad j = 1, \ldots, n_i, \\ \theta_i &\sim \mathcal{N}(\mu, \tau^2), \quad i = 1, \ldots, k, \\ \mu &\sim \mathcal{N}(\mu_0, \sigma_\mu^2), \\ \sigma^2 &\sim \mathcal{IG}(a_1, b_1), \quad \tau^2 \sim \mathcal{IG}(a_2, b_2), \quad \sigma_\mu^2 \sim \mathcal{IG}(a_3, b_3). \end{aligned} \tag{7.6}$$

Now, if we proceed as before and write down the joint distribution from this hierarchy, we can derive the set of full conditionals

$$\begin{aligned} \theta_i &\sim \mathcal{N}\left(\frac{\sigma^2}{\sigma^2 + n_i\tau^2}\mu + \frac{n_i\tau^2}{\sigma^2 + n_i\tau^2}\bar{X}_i,\ \frac{\sigma^2\tau^2}{\sigma^2 + n_i\tau^2}\right), \quad i = 1, \ldots, k, \\ \mu &\sim \mathcal{N}\left(\frac{\tau^2}{\tau^2 + k\sigma_\mu^2}\mu_0 + \frac{k\sigma_\mu^2}{\tau^2 + k\sigma_\mu^2}\bar{\theta},\ \frac{\sigma_\mu^2\tau^2}{\tau^2 + k\sigma_\mu^2}\right), \end{aligned}$$

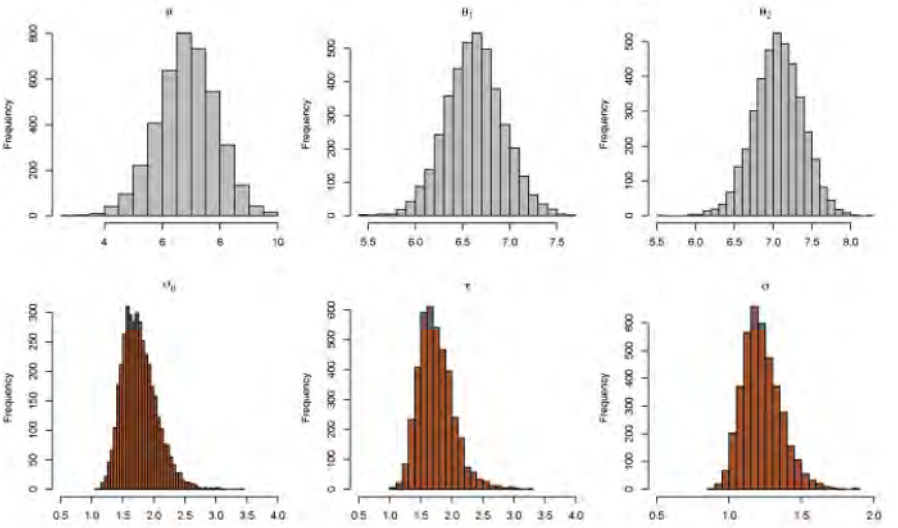


**Fig. 7.3.** Histograms of marginal posterior distributions from the Gibbs sampler of Example 7.5 based on 5000 iterations. The top row gives histograms for the underlying mean $\mu$ and the means, $\theta_1$ and $\theta_2$, for the girls' and boys' energy. The bottom row corresponds to the standard deviations.

$$
\begin{aligned}
(7.7) \qquad \sigma^2 &\sim \mathcal{IG}\left(n/2 + a_1,\ (1/2)\sum_{ij}(X_{ij} - \theta_i)^2 + b_1\right), \\
\tau^2 &\sim \mathcal{IG}\left(k/2 + a_2,\ (1/2)\sum_{i}(\theta_i - \mu)^2 + b_2\right), \\
\sigma_\mu^2 &\sim \mathcal{IG}\left(1/2 + a_3,\ (1/2)(\mu - \mu_0)^2 + b_3\right),
\end{aligned}
$$

where $n = \sum_i n_i$ and $\bar{\theta} = \sum_i n_i\theta_i/n$.

Expanding on the study in Example 7.3, the dataset `Energy` also contains data on the energy intake of boys. Model (7.6) applies (with $k = 2$) to the simultaneous analysis of the energy intakes of girls and boys. The outcome of the Gibbs sampler based on the conditionals in (7.7) is summarized in Figure 7.3. ◀

**Exercise 7.6** In the setting of Example 7.5:

a. Derive the full conditional distributions in (7.7).
b. Implement this Gibbs sampler in `R` to reproduce the histograms in Figure 7.3.
c. A variation on the model (7.6) is to give $\mu$ a flat prior, which is equivalent to setting $\sigma_\mu^2 = \infty$ in (7.6). Construct the full conditionals for this model and modify the previous `R` code to compare both models on the `Energy` data.

## 7.4 Missing data and latent variables

Starting with the two-stage Gibbs sampler, working on a joint distribution $f(x, y)$, there seems to be a major difference with the Metropolis–Hastings algorithm that works with a single distribution or, in other words, generates all components of $(x, y)$ at once. This difference in the target is illusory in that once given $f(x, y)$ we can use either the relevant Gibbs sampler or a generic Metropolis–Hastings algorithm, while if given a marginal density $f_X(x)$, we can construct (or *complete* $f_X(x)$ into) a joint density $f(x, y)$ to aid in simulation, where the second variable $Y$ is then an *auxiliary variable* that may not be directly relevant from a statistical point of view. There are many settings where a natural completion of $f_X(x)$ into $f(x, y)$ does exist and in fact this can lead to an effective Gibbs sampler.[2]

These considerations bring us back into the realm of *missing-data models*, as described in Section 5.4.2, where the representation (5.9)

$$g(x|\theta) = \int_{\mathcal{Z}} f(x, z|\theta)\, \mathrm{d}z$$

was introduced. As discussed in Chapter 5, $g(x|\theta)$ is the density of the observations (that is, the likelihood), and the right side represents the completion joint density. The density $f$ is arbitrary and can be chosen so that the full conditionals of $f$ are easy to simulate from and the Gibbs algorithm (Algorithm 8) is implemented on $f$ instead of $g$ and the corresponding full conditional of $\theta$ given $(x, z)$.

Depending on the field, such representations go by different names. From a mathematical perspective, (5.9) is a mixture model. In statistics, we most often use the name of missing-data models, while econometricians prefer the use of *latent variable* models, maybe because of the related feeling of *deus ex machina* operating behind the scenes! If we factor $f(x, z|\theta) = f(x|z, \theta)h(z|\theta)$, then (5.9) becomes

$$g(x|\theta) = \int_{\mathcal{Z}} f(x|z, \theta)h(z|\theta)\, \mathrm{d}z,$$

and $h(z|\theta)$, the marginal distribution of the missing data $z$, is clearly a mixing distribution.

In a general missing-data setting,

$$g(x) = \int_{\mathcal{Z}} f(x, z)\, \mathrm{d}z$$

[2] It is obviously always the case that any given density $f_X(x)$ can be artificially completed into a joint density $f(x, y)$, as demonstrated with the slice sampler at the end of this section.

for $p \geq 2$, we write $y = (x, z) = (y_1, \ldots, y_p)$ and denote the conditional densities of $f(y) = f(y_1, \ldots, y_p)$ by

$$\begin{aligned} Y_1|y_2, \ldots, y_p &\sim f_1(y_1|y_2, \ldots, y_p), \\ Y_2|y_1, y_3, \ldots, y_p &\sim f_2(y_2|y_1, y_3, \ldots, y_p), \\ &\vdots \\ Y_p|y_1, \ldots, y_{p-1} &\sim f_p(y_p|y_1, \ldots, y_{p-1}). \end{aligned}$$

Then, applying a multistage Gibbs sampler as in Algorithm 8 to those full conditionals and assuming they all can be simulated leads to a Markov $(Y^{(t)})_t$ that converges to $f$ and therefore a subchain $(X^{(t)})_t$ that converges to $g$.

**Example 7.6.** In Examples 5.13 and 5.14, we treated a censored-data model as a missing-data model. We identify $g(x|\theta)$ with the likelihood function

$$g(x|\theta) = L(\theta|x) \propto \prod_{i=1}^{m} e^{-(x_i-\theta)^2/2},$$

and

$$f(x, z|\theta) = L(\theta|x, z) \propto \prod_{i=1}^{m} e^{-(x_i-\theta)^2/2} \prod_{i=m+1}^{n} e^{-(z_i-\theta)^2/2}$$

is the complete-data likelihood. Given a prior distribution $\pi(\theta)$ on $\theta$, we can then create a Gibbs sampler that iterates between the conditional distributions

$$\pi(\theta|x, z) \quad \text{and} \quad f(z|x, \theta)$$

and will have stationary distribution $\pi(\theta, z|x)$, the posterior distribution of $(\theta, z)$.

Taking a flat prior $\pi(\theta) = 1$, the conditional distribution of $\theta|x, z$ is given by

$$\theta|x, z \sim \mathcal{N}\left(\frac{m\bar{x} + (n-m)\bar{z}}{n}, \frac{1}{n}\right),$$

while the conditional distribution of $Z|x, \theta$ is the product of the truncated normals

$$Z_i|x, \theta \sim \varphi(z-\theta) \Big/ \{1 - \Phi(a-\theta)\},$$

as each $Z_i$ must be greater than the truncation point $a$. Generating values of $Z$ can be done via the `R` function `rtrun` from the package `bayesm` (see Exercises 7.21 and 7.7). The outcome of the Gibbs sampler, whose `R` core can be written as

```
> for(i in 2:Nsim){
>   zbar[i]=mean(rtrun(mean=rep(that[i-1],n-m),
+   sigma=rep(1,n-m),a=rep(a,n-m),b=rep(Inf,n-m)))
>   that[i]=rnorm(1,(m/n)*xbar+(1-m/n)*zbar[i],sqrt(1/n))
>   }
```

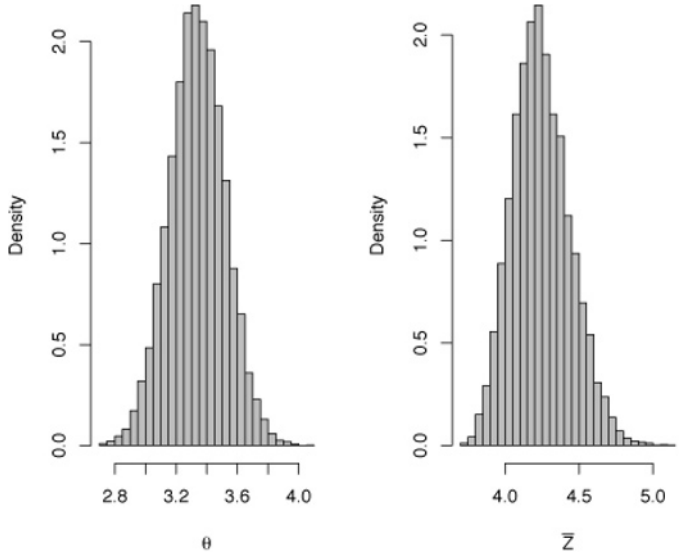


**Fig. 7.4.** Histograms of the posterior distributions of $\theta$ and $\bar{Z}$ from Example 7.6. The truncation point on the $Z_i$'s is $a = 3.5$.

is summarized in Figure 7.4 using the posterior distributions of $\theta$ and $\bar{Z}$. ◀

**Exercise 7.7** Referring to Example 7.6:

a. Show that, as a function of $\theta$, the complete data likelihood is proportional to the density of $\mathcal{N}(\{m\bar{x} + (n-m)\bar{z}\}/n, 1/n)$.
b. Complete the `R` code above into a Gibbs sampler that estimates the posterior distribution of $\theta$.

**Example 7.7.** Recall the multinomial model of Example 5.16,

$$\mathcal{M}\left(n; \frac{1}{2} + \frac{\theta}{4}, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{\theta}{4}\right).$$

where we estimated $\theta$ using either EM or MCEM steps, introducing the latent variable $Z$ with the demarginalization

$$(z, x_1 - z, x_2, x_3, x_4) \sim \mathcal{M}\left(n; \frac{1}{2}, \frac{\theta}{4}, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{\theta}{4}\right).$$

If we use a uniform prior on $\theta$, the full conditionals can be recovered as

$$\theta \sim \mathcal{B}e(z + x_4 + 1, x_2 + x_3 + 1) \text{ and } z \sim \mathcal{B}\text{in}\left(x_1, \frac{\theta}{2+\theta}\right),$$

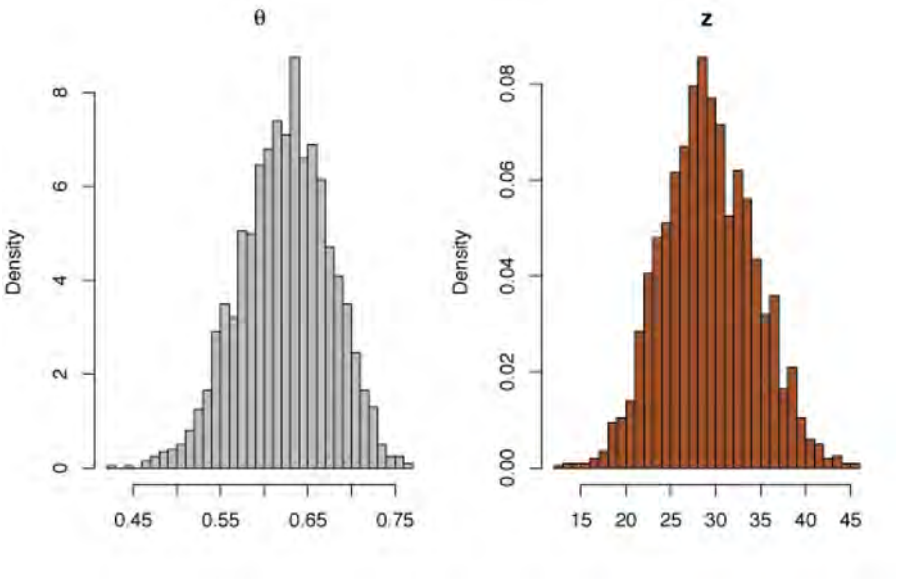


**Fig. 7.5.** Histograms of marginal distributions from the Gibbs sampler of Example 7.7. The main interest is in the marginal distribution of $\theta$.

leading to the Gibbs sampler

```
> x=c(125,18,20,34)              #data
> theta=z=rep(.5,Nsim)           #init chain
> for (j in 2:Nsim){
>   theta[j]=rbeta(1,z[j-1]+x[4]+1,x[2]+x[3]+1)
>   z[j]=rbinom(1,x{1},(theta[j]/(2+theta[j])))
>   }
```

whose output is summarized in Figure 7.5. ◀

This example shows a case where both EM and the Gibbs sampler apply. As usual, the Bayesian approach allows a more complete inference that includes confidence intervals.

**Exercise 7.8** In the setting of Example 7.7:

a. Construct a 95% confidence interval for $\theta$ based on the outcome of the Gibbs sampler, and verify whether or not the EM solution belongs to this interval.
b. The Gibbs sampler above used a uniform prior on $\theta$; that is, $\theta \sim \mathcal{B}e(a, b)$ with $a = b = 1$. Write a Gibbs sampler for general $a$ and $b$, and, for a range of $a$ and $b$, compare the Gibbs estimates of $\theta$ with the EM answer. What can you conclude about sensitivity to the prior?

**Example 7.8.** A generalization of the model of Example 7.7 is the model

$$X \sim \mathcal{M}_5\left(n; a_1\theta_1 + b_1, a_2\theta_1 + b_2, a_3\theta_2 + b_3, a_4\theta_2 + b_4, c(1-\theta_1-\theta_2)\right), \tag{7.8}$$

with $0 \le a_1 + a_2 = a_3 + a_4 = 1 - \sum_{i=1}^4 b_i = c \le 1$, where the $a_i, b_i \ge 0$ are known based on genetic considerations, as in Table 7.1 describing the probabilities of the four blood types as functions of genotype probabilities, because of allele dominance. Our interest is in estimating the allele frequencies $p_A, p_B$, and $p_O$ (which sum to 1).

We can then augment the data with $\mathbf{Z} = (Z_1, Z_2, Z_3, Z_4)$ as

$$X_1 = Z_1 + Z_2,\ X_2 = Z_3 + Z_4,\ X_3 = Z_5 + Z_6,\ X_4 = Z_7 + Z_8,$$

which demarginalizes the model to allow us to sample from

$$Y \sim \mathcal{M}_9\left(n; a_1\theta_1, b_1, a_2\theta_1, b_2, a_3\theta_2, b_3, a_4\theta_2, b_4, c(1-\theta_1-\theta_2)\right),$$

with $Y = (Z_1, X_1 - Z_1, Z_2, X_2 - Z_2, Z_3, X_3 - Z_3, Z_4, X_4 - Z_4, X_5)$. (See Exercise 7.23 for an alternate solution.) A natural prior distribution on $(\theta_1, \theta_2)$ is the Dirichlet prior $\mathcal{D}(\alpha_1, \alpha_2, \alpha_3)$,

$$\pi(\theta_1, \theta_2) \propto \theta_1^{\alpha_1 - 1}\theta_2^{\alpha_2 - 1}(1-\theta_1-\theta_2)^{\alpha_3 - 1},$$

which leads to the full conditionals

$$\begin{aligned}
(\theta_1, \theta_2, 1-\theta_1-\theta_2)|x, \mathbf{z} &\sim \mathcal{D}(z_1 + z_2 + \alpha_1, z_3 + z_4 + \alpha_2, x_5 + \alpha_3), \\
Z_i|x, \theta_1, \theta_2 &\sim \mathcal{B}\left(x_i, \frac{a_i\theta_1}{a_i\theta_1 + b_i}\right) \qquad (i = 1, 3), \\
Z_i|x, \theta_1, \theta_2 &\sim \mathcal{B}\left(x_i, \frac{a_i\theta_2}{a_i\theta_2 + b_i}\right) \qquad (i = 5, 7),
\end{aligned} \tag{7.9}$$

which can all easily be simulated and thus included within a Gibbs sampler. Figure 7.6 shows the distributions of chains produced by such a sampler. ◀

**Table 7.1.** Observed genotype frequencies on blood type data. The effect of a dominant allele creates a missing-data problem.

| Genotype | Probability | Observed | Probability | Frequency |
|---|---|---|---|---|
| AA | $p_A^2$ | A | $p_A^2 + 2p_Ap_O$ | $n_A = 186$ |
| AO | $2p_Ap_O$ | | | |
| BB | $p_B^2$ | B | $p_B^2 + 2p_Bp_O$ | $n_B = 38$ |
| BO | $2p_Bp_O$ | | | |
| AB | $2p_Ap_B$ | AB | $p_Ap_B$ | $n_{AB} = 13$ |
| OO | $p_O^2$ | O | $p_O^2$ | $n_O = 284$ |

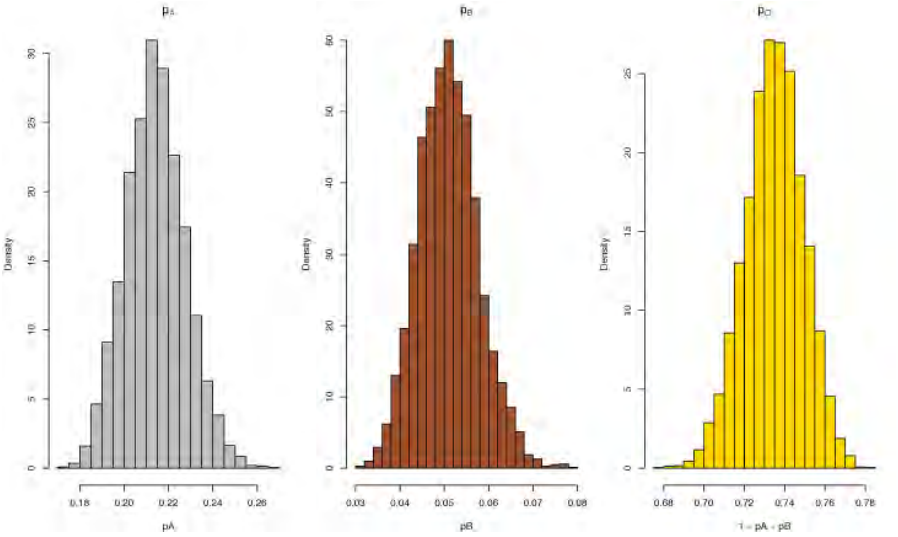


**Fig. 7.6.** Histograms of marginal distributions of the genotype probabilities from the Gibbs sampler of Example 7.8.

**Exercise 7.9** For the data in Table 7.1, modeled with (7.8):

a. Verify that the observed data likelihood is proportional to

$$(p_A^2 + 2p_Ap_O)^{n_A}(p_B^2 + 2p_Bp_O)^{n_B}(p_Ap_B)^{n_{AB}}(p_O^2)^{n_O}.$$

b. With missing data $Z_A$ and $Z_B$, verify that the complete-data likelihood is proportional to

$$(p_A^2)^{Z_A}(2p_Ap_O)^{n_A-Z_A}(p_B^2)^{Z_B}(2p_Bp_O)^{n_B-Z_B}(p_Ap_B)^{n_{AB}}(p_O^2)^{n_O}.$$

c. Write a Gibbs sampler to estimate $p_A$ and $p_B$.

Finite mixture models, which we have seen in some detail in Chapter 5 (Example 5.2) and Chapter 6 (Example 6.5), are obviously a candidate for demarginalization through latent variables; that is, as a special case of a mixture (!). As already described in Example 5.12, given a sample $(x_1, \ldots, x_n)$ from a mixture distribution

$$\sum_{j=1}^{k} p_j \, f(x|\xi_j) \, ,$$

where $\sum_j p_j = 1$ and $f(\cdot|\xi_j)$ is a parameterized density with unknown parameter $\xi_j$, we can associate with every observation $x_i$ a latent variable

$z_i \in \{1, \ldots, k\}$ that indicates which component of the mixture is associated with $x_i$. The corresponding completion of the mixture model above is then

$$Z_i \sim \mathcal{M}_k(1; p_1, \ldots, p_k), \qquad x_i | z_i \sim f(x|\xi_{z_i}).$$

Thus, considering $y_i = (x_i, z_i)$ (instead of $x_i$) entirely eliminates the mixture structure since the likelihood of the completed model is

$$\ell(p, \xi | y_1, \ldots, y_n) \propto \prod_{i=1}^{n} p_{z_i} \, f(x_i|\xi_{z_i}) = \prod_{j=1}^{k} \prod_{i; z_i = j} p_j \, f(x_i|\xi_j).$$

One may wonder why the completion is useful in this setting since the observed likelihood can be computed in closed form, as shown for instance in Figure 5.2, which represents a mixture likelihood on a grid of pixels as in Example 6.5, where we produced a random walk Metropolis–Hastings algorithm. As in the EM algorithm of Examples 5.12 and 5.13, using the latent indicator variables produces a usually efficient simulation algorithm that quickly focuses on the mode(s) of the posterior distribution.

The two steps of the Gibbs sampler are then associated with the full conditional posteriors

$$P(Z_i = j | \mathbf{x}, \xi) \propto p_j \, f(x_i|\xi_j) \qquad (i = 1, \ldots, n, \, j = 1, \ldots, k)$$

and

$$\xi_j | \mathbf{y} \sim \pi\left(\xi \middle| \frac{\lambda_j \alpha_j + n_j \overline{x}_j}{\lambda_j + n_j}, \lambda_j + n_j\right),$$
$$p \sim \mathcal{D}_k(\gamma_1 + n_1, \ldots, \gamma_k + n_k),$$

where

$$n_j = \sum_{i=1}^{n} \mathbb{I}_{z_i = j}, \qquad n_j \overline{x}_j = \sum_{i=1}^{n} \mathbb{I}_{z_i = j} x_i.$$

In this two-step Gibbs sampler, the generation from the posterior associated with the complete likelihood is not detailed, as it will vary depending on the sampling model and the prior used. In the standard situation relying on an exponential family for $f(\cdot|\xi)$ and a conjugate prior on $\xi$, this generation is obviously straightforward.

**Example 7.9.** As an illustration, consider the same setting as in Example 5.12, namely a normal mixture with two components with equal known variance and fixed weights,

$$p\mathcal{N}(\mu_1, \sigma^2) + (1 - p)\mathcal{N}(\mu_2, \sigma^2)\,.$$

We assume in addition a normal $\mathcal{N}(0, v^2\sigma^2)$ prior distribution, with $v^2$ known, on both means $\mu_1$ and $\mu_2$. The latent variables $z_i$ are the same as in Example 5.12, namely

$$P(Z_i = 1) = 1 - P(Z_i = 2) = p \qquad \text{and} \qquad X_i|Z_i = k \sim \mathcal{N}(\mu_k, \sigma^2)\,.$$

The completed distribution is then

$$\pi(\mu_1, \mu_2, \mathbf{z}|\mathbf{x}) \propto \exp\left\{-(\mu_1^2 + \mu_2^2)/v^2\sigma^2\right\} \times \prod_{i:z_i=1} p \exp\left\{-\frac{(x_i - \mu_1)^2}{2\sigma^2}\right\} \prod_{i:z_i=2} (1-p) \exp\left\{-\frac{(x_i - \mu_2)^2}{2\sigma^2}\right\},$$

for which the full conditionals of the $\mu_j$'s are easily derived (Exercise 7.10). Figure 7.7 illustrates the behavior of the corresponding Gibbs sampler using a simulated dataset $\mathbf{x}$ of $500$ points from the $.7\mathcal{N}(0,1)+.3\mathcal{N}(2.7,1)$ distribution. This picture plots the MCMC sample after $15,000$ iterations on top of the log-posterior surface. This simulation is in fact in clear agreement with the posterior surface. Although it may appear to be too concentrated around one mode, you must account for the fact that the second mode represented on this graph is much lower since there is a difference of at least $50$ in log-posterior values. ◀

**Exercise 7.10** Using the completed joint distribution in Example 7.9:

a. Show that the conditional distributions are $j = 1, 2$

$$\mu_j|\mathbf{x}, \mathbf{z} \sim \mathcal{N}\left(\frac{v^2}{n_j v^2 + 1} \sum_{i;z_i=j} x_i,\ \frac{\sigma^2 v^2}{n_j v^2 + 1}\right),$$

where $n_j$ denotes the number of $z_i$'s equal to $j$ and

$$P(Z_i = j|x_i, \mu_1, \mu_2) = \frac{p \exp\left\{-\frac{(x_i-\mu_j)^2}{2\sigma^2}\right\}}{p \exp\left\{-\frac{(x_i-\mu_1)^2}{2\sigma^2}\right\} + (1-p)\left\{-\frac{(x_i-\mu_2)^2}{2\sigma^2}\right\}}\,.$$

b. Write the `R` code to reproduce Figure 7.7.

c. For $\sigma = 1$, investigate the convergence of the Gibbs sampler for various combinations of the true values of $(\mu_1, \mu_2, p)$. In particular, you should find that if the $\mu_i$s are too separated and $p = 0.5$, the Gibbs sampler may concentrate in one mode even though the modal likelihoods are similar.

As a last "example" of a latent variable Gibbs sampler, we look at the *slice sampler*, which appears more like a generic type of demarginalization.[3] (See Neal, 2003, for a comprehensive treatment.) Given a density of interest $f_X(x)$, we can always represent it as the marginal density of the joint density

[3] In fact, we can alternatively consider the Gibbs sampler as being derived from the slice sampler; see Robert and Casella (2004, Chapter 8).

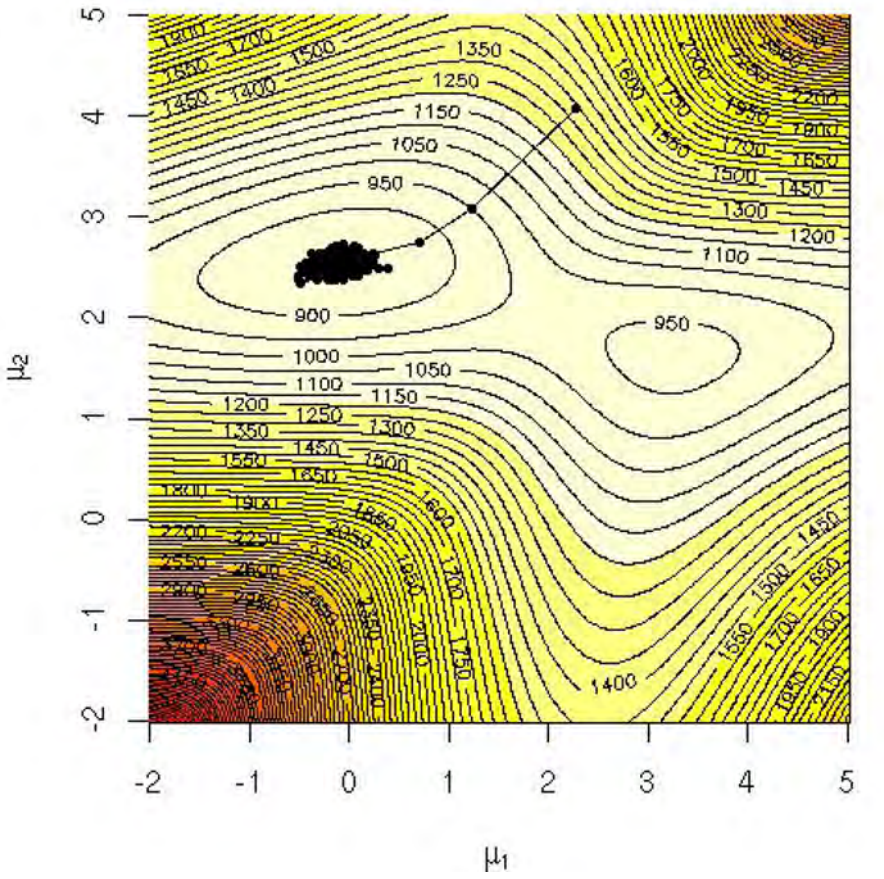


**Fig. 7.7.** Gibbs sample of 5000 points for the mixture posterior against the log-posterior surface.

$$f(x,u) = \mathbb{I}\{0 < u < f_X(x)\}$$

since integrating the above in $u$ returns $f_X$. The associated conditional densities are

$$f_{X|U}(x|u) = \frac{\mathbb{I}\{0 < u < f_X(x)\}}{\int \mathbb{I}\{0 < u < f_X(x)\}\,\mathrm{d}x}\,, f_{U|X}(u|x) = \frac{\mathbb{I}\{0 < u < f_X(x)\}}{\int \mathbb{I}\{0 < u < f_X(x)\}\,\mathrm{d}u}\,,$$

which means they are both uniform. Those two conditionals then define the slice sampler as the associated Gibbs sampler.

**Algorithm 9 2D slice sampler**
At iteration $t$, simulate

1. $U^{(t+1)} \sim \mathcal{U}_{[0,f(x^{(t)})]}$;
2. $X^{(t+1)} \sim \mathcal{U}_{A^{(t+1)}}$, with

$$A^{(t+1)} = \{x : \ f(x) \geq u^{(t+1)}\}.$$

The appeal of this algorithm is that it formally applies to any density known up to a multiplicative constant with no restriction on its shape or dimension. Obviously, its implementation may be hindered by the uniform simulation over the set $A^{(t)}$.

**Example 7.10.** Consider the density $f(x) = \frac{1}{2}e^{-\sqrt{x}}$ defined for $x > 0$. While it can be directly simulated, it also yields easily to the slice sampler. Indeed, applying the formulas above, we have

$$U|x \sim \mathcal{U}\left(0, \frac{1}{2}e^{-\sqrt{x}}\right), \quad X|u \sim \mathcal{U}\left(0, [\log(2u)]^2\right).$$

We implement the sampler to generate 5000 variates and plot them along with the density in Figure 7.8, which shows that the agreement is very good. The right panel does show some strong autocorrelations, which is typical of the slice sampler. ◀

**Exercise 7.11** Referring to Example 7.10 and the density $f_X(x) = (1/2)\exp(-\sqrt{x})$:

a. Verify that the conditional distributions are

$$U|x \sim \mathcal{U}\left(0, (1/2)\exp(-\sqrt{x})\right) \text{ and } X|u \sim \mathcal{U}\left(0, [\log(2u)]^2\right),$$

and implement a Gibbs sampler to generate random variables from $f_X(x)$.

b. Make the transformation $Y = \sqrt{X}$ and show that $Y \sim \mathcal{G}(3/2, 1)$. Use this fact to simulate directly $X$. Compare this algorithm with the slice sampler.

There is an obvious extension to the 2D slice sampler above, akin to the multistage extension to the two-stage Gibbs sampler. If the target density is written as a product of functions,

$$f(x) = \prod_{i=1}^{n} g_i(x),$$

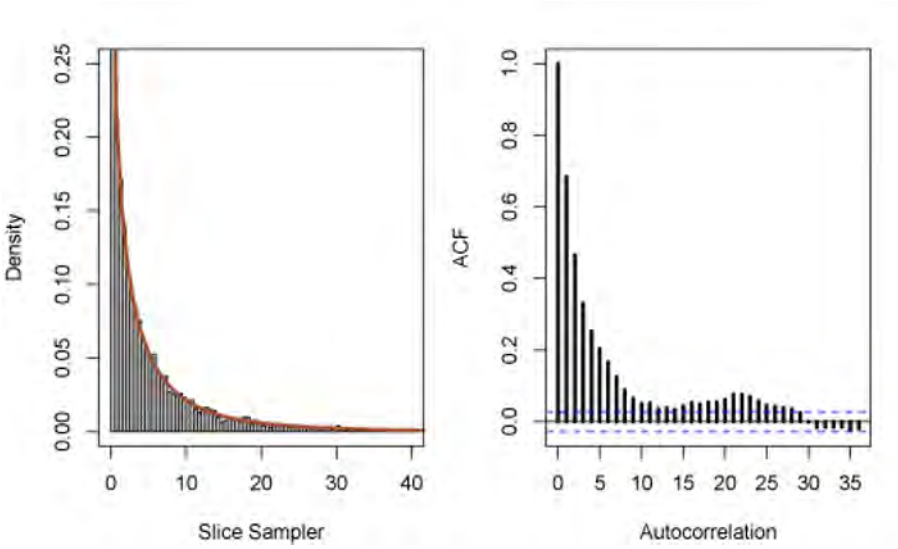


**Fig. 7.8.** A slice sampler histogram and target density for Example 7.10 using 5000 iterations. The left panel is the histogram with the true density overlaid, and the right panel shows the autocorrelation function.

as for instance in the case of a posterior distribution associated with a sample of $n$ observations (where the $g_i$'s are then the componentwise densities), an associated completion is

$$f(x, u_1, \ldots, u_n) = \prod_{i=1}^{n} \mathbb{I}\{0 < u_i < g_i(x)\},$$

which leads to a slice sampler with $(n+1)$ steps, $X^{(t)}$ then being uniformly generated over the set

$$A^{(t)} = \bigcap_{i=1}^{n} \left\{x \,:\, g_i(x) > u_i^{(t)}\right\}.$$

**Example 7.11.** Recall logistic regression, which we first saw in Example 4.11 and fit with a Metropolis–Hastings algorithm in Exercise 6.13. The model is

$$Y_i \sim \text{Bernoulli}(p(x_i)), \quad p(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)},$$

where $p(x)$ is the success probability and $x$ is a unidimensional covariate. The likelihood associated with a sample $(\mathbf{y}, \mathbf{x}) = (y_1, x_1), \ldots, (y_n, x_n)$ is

$$L(\alpha, \beta|\mathbf{y}) \propto \prod_{i=1}^{n} \left(\frac{e^{\alpha+\beta x_i}}{1 + e^{\alpha+\beta x_i}}\right)^{y_i} \left(\frac{1}{1 + e^{\alpha+\beta x_i}}\right)^{1-y_i}.$$

Using a flat prior on $(a, b)$, the posterior distribution can be associated with a slice sampler based on uniform

$$U_i \sim \mathcal{U}\left(0, \frac{e^{y_i(\alpha+\beta x_i)}}{1+e^{\alpha+\beta x_i}}\right)$$

variables. Generating a uniform distribution over the set

$$\left\{(a,b)\ :\ y_i(a+bx_i) > \log\frac{u_i}{1-u_i}\right\}$$

being rather unwieldy, we can further decompose the uniform simulation by consecutively simulating

$$a^{(t)} \sim \mathcal{U}\left(\max_{i;y_i=1} \log\frac{u_i^{(t)}}{1-u_i^{(t)}} - b^{(t-1)}x_i, \min_{i;y_i=0} \log\frac{1-u_i^{(t)}}{u_i^{(t)}} - b^{(t-1)}x_i\right)$$

and

$$b^{(t)} \sim \mathcal{U}\left(\max_{i;y_i=1}\left[\log\frac{u_i^{(t)}}{1-u_i^{(t)}} - a^{(t)}\right]/x_i, \min_{i;y_i=0}\left[\log\frac{1-u_i^{(t)}}{u_i^{(t)}} - a^{(t)}\right]/x_i\right),$$

if we assume without loss of generality that all $x_i$'s are positive. However, running the corresponding slice sampler on the `challenger` dataset described in Exercise 6.13 exhibits a random walk behavior on the chain $(a^{(t)}, b^{(t)})_t$, as shown in Figure 7.9. We therefore introduce instead normal $\mathcal{N}(0, \sigma^2)$ priors on both $a$ and $b$. The modification on the slice sampler is minimal in that both uniform distributions above are replaced with truncated normals $\mathcal{N}(0, \sigma^2)$, the truncation intervals being those used above. The core of the `R` code is then

```
> for (t in 2:Nsim){
+   uni=runif(n)*exp(y*(a[t-1]+b[t-1]*x))/
+               (1+exp(a[t-1]+b[t-1]*x))
+   mina=max(log(uni[y==1]/(1-uni[y==1]))-b[t-1]*x[y==1])
+   maxa=min(-log(uni[y==0]/(1-uni[y==0]))-b[t-1]*x[y==0])
+   a[t]=rtrun(0,sigmaa,mina,maxa)
+   minb=max((log(uni[y==1]/(1-uni[y==1]))-a[t])/x[y==1])
+   maxb=min((-log(uni[y==0]/(1-uni[y==0]))-a[t])/x[y==0])
+   b[t]=rtrun(0,sigmab,minb,maxb)
+   }
```

with `sigmaa` equal to $5$ and `sigmab` equal to $5$ divided by the standard deviation of the $x_i$'s. ◀

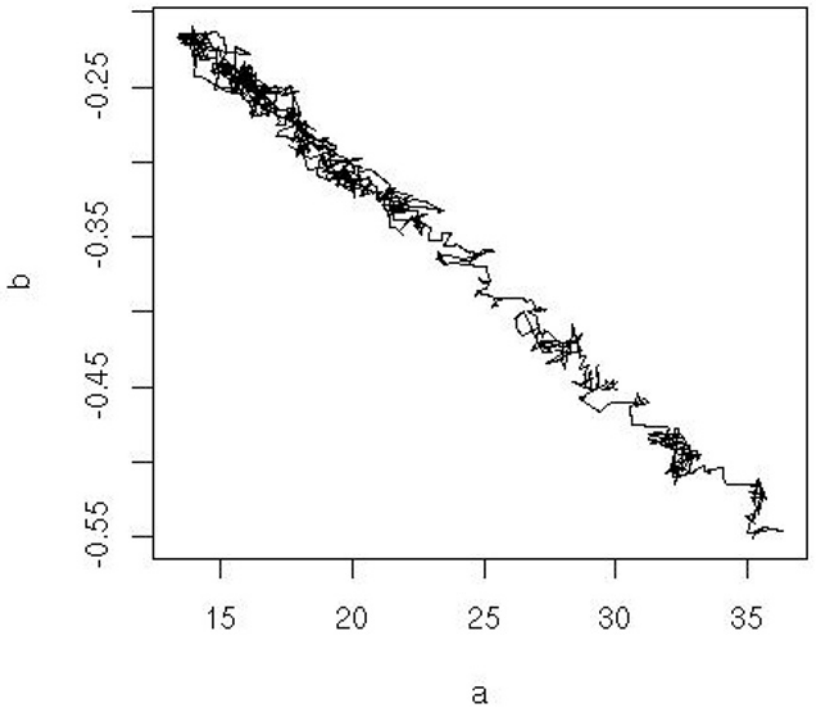


**Fig. 7.9.** Evolution of the chain $(a^{(t)}, b^{(t)})_t$ along $10^3$ final iterations of a slice sampler for the `challenger` dataset under a flat prior.

## 7.5 Hierarchical structures

We have seen the multistage Gibbs sampler applied to a number of examples, most arising from missing-data structures. However, it is equally well-suited to sample in a straightforward way from any hierarchical model.

A *hierarchical model* is defined by a sequence of conditional distributions as, for instance, in the two-level generic hierarchy

$$\begin{aligned} X_i &\sim f_i(x|\theta), \quad i = 1, \ldots, n\,, \quad \theta = (\theta_1, \ldots, \theta_p)\,, \\ \theta_j &\sim \pi_j(\theta|\gamma), \quad j = 1, \ldots, p\,, \quad \gamma = (\gamma_1, \ldots, \gamma_s)\,, \\ \gamma_k &\sim g(\gamma), \quad k = 1, \ldots, s. \end{aligned}$$

The joint distribution from this hierarchy is

$$\prod_{i=1}^{n} f_i(x_i|\theta) \prod_{j=1}^{p} \pi_j(\theta_j|\gamma) \prod_{k=1}^{s} g(\gamma_k)\,.$$

Assuming that the $x_i$'s are observations, the corresponding posterior distribution on $(\theta, \gamma)$ is associated with the full posterior conditionals

$$\theta_j \propto \pi_j(\theta_j|\gamma) \prod_{i=1}^{n} f_i(x_i|\theta), \quad j = 1, \ldots, p\,,$$
$$\gamma_k \propto g(\gamma_k) \prod_{j=1}^{p} \pi_j(\theta_j|\gamma), \quad k = 1, \ldots, s\,.$$

In standard hierarchies, these densities are straightforward to simulate from and are therefore naturally associated with a Gibbs sampler. In more complex hierarchies, we might need to use more sophisticated methods, such as a Metropolis–Hastings step or another slice sampler, to sample from the conditionals (as explained in Section 7.6.3). However, our main message here is that the full conditionals are quite easy to write down given the hierarchical specification, while they considerably reduce the dimension of the random variables to simulate at each step.

ϟ When a full conditional in a Gibbs sampler cannot be simulated directly, it is sufficient to run instead a single step of any MCMC algorithm associated with this full conditional. The theoretical validation is the same as with any MCMC sampler. In the event a slice sampler is used for this purpose, the auxiliary variable is simply added to the vector of parameters.

**Example 7.12.** A benchmark hierarchical example in the Gibbs sampling literature describes multiple failures of ten pumps in a nuclear plant, with the data given in Table 7.2. The modeling is based on the assumption that the number of

**Table 7.2.** Number of failures and times of observation of ten pumps in a nuclear plant (*source*: Gaver and O'Muircheartaigh, 1987).

| Pump | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Failures | 5 | 1 | 5 | 14 | 3 | 19 | 1 | 1 | 4 | 22 |
| Time | 94.32 | 15.72 | 62.88 | 125.76 | 5.24 | 31.44 | 1.05 | 1.05 | 2.10 | 10.48 |

failures of the $i$th pump follows a Poisson process with parameter $\lambda_i$ $(1 \leq i \leq 10)$. For an observation time $t_i$, the number of failures $X_i$ is thus a Poisson $\mathcal{P}(\lambda_i t_i)$ random variable. The standard prior distributions are gamma distributions, which lead to the hierarchical model

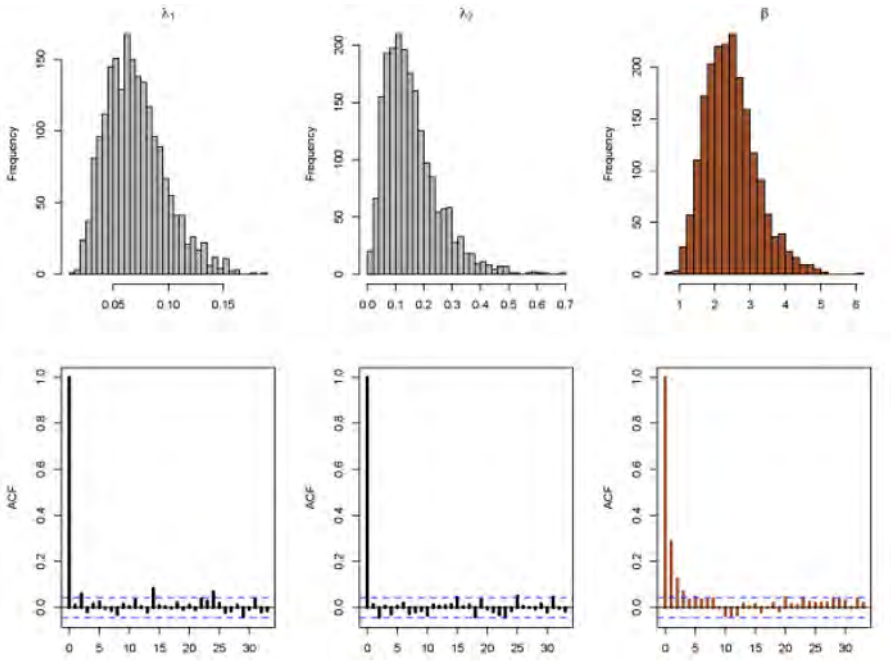


**Fig. 7.10.** Histograms of marginal distributions of $\lambda_1$, $\lambda_2$, and $\beta$ from the pump failure data of Example 7.12. The corresponding bottom panels are autocorrelation plots. The hyperparameter values are $\alpha = 1.8$, $\gamma = 0.01$, and $\delta = 1$.

$$\begin{aligned} X_i &\sim \mathcal{P}(\lambda_i t_i), \quad i = 1, \ldots 10, \\ \lambda_i &\sim \mathcal{G}(\alpha, \beta), \quad i = 1, \ldots 10, \\ \beta &\sim \mathcal{G}(\gamma, \delta). \end{aligned}$$

The joint distribution is thus

$$\begin{aligned} &\pi(\lambda_1, \ldots, \lambda_{10}, \beta | t_1, \ldots, t_{10}, p_1, \ldots, p_{10}) \\ &\propto \prod_{i=1}^{10} \left\{ (\lambda_i t_i)^{x_i} \; e^{-\lambda_i t_i} \; \lambda_i^{\alpha-1} e^{-\beta \lambda_i} \right\} \beta^{10\alpha} \beta^{\gamma-1} e^{-\delta\beta} \\ &\propto \prod_{i=1}^{10} \left\{ \lambda_i^{x_i+\alpha-1} \; e^{-(t_i+\beta)\lambda_i} \right\} \beta^{10\alpha+\gamma-1} e^{-\delta\beta}, \end{aligned}$$

leading to the full conditional distributions

$$\lambda_i | \beta, t_i, x_i \sim \mathcal{G}(x_i + \alpha, t_i + \beta), \quad i = 1, \ldots 10,$$

$$\beta | \lambda_1, \ldots, \lambda_{10} \sim \mathcal{G}\left( \gamma + 10\alpha, \delta + \sum_{i=1}^{10} \lambda_i \right).$$

The associated Gibbs sampler is quite straightforward, with core R code

```
> for(i in 2:Nsim){
+  for(j in 1:nx)
+   lambda[i,j]=rgamma(1,sh=xdata[j]+alpha,ra=Time[j]+beta[i-1])
+  beta[i]=rgamma(1,sh=gamma+nx*alpha,ra=delta+sum(lambda[i,]))}
```

The result of a run over $5000$ iterations is shown in Figure 7.10. ◀

**Exercise 7.12** One reason for collecting the pump failure data is to identify which pumps are more reliable.

a. Run the Gibbs sampler for the pump failure data and get $95\%$ posterior credible intervals for the parameters $\lambda_i$.
b. Based on the analysis, can you identify any pumps that are more or less reliable than the others?
c. How does your answer in b. change as the hyperparameter values are varied?

## 7.6 Other considerations

In this last section, we look at a few issues that could arise in the implementation of a Gibbs sampler.

### 7.6.1 Reparameterization

Many factors contribute to the convergence properties of a Gibbs sampler. For example, convergence performance may be greatly affected by the choice of the coordinates (or, in other words, the parameterization). If the covariance matrix $\Sigma$ of the target has a wide range of eigenvalues, the Gibbs sampler may be very slow to explore the entire range of the support of the target.

**Example 7.13.** Recall Example 7.1, where we saw a first Gibbs sampler for the bivariate normal in (7.1). For that bivariate normal distribution, Figure 7.11 shows the autocorrelation for $\rho = .3, .6, .9$. The higher correlation results in a sampler that will have more trouble exploring the entire space and thus require more iterations. It is also interesting to note that no matter what is the value of $\rho$, $X+Y$ and $X-Y$ are independent, and thus changing coordinates from $(x, y)$ to $(x+y, x-y)$ would lead to an immediately converging Gibbs algorithm. ◀

**Exercise 7.13** For the bivariate normal distribution (7.1):

a. prove that $X+Y$ and $X-Y$ are independent.

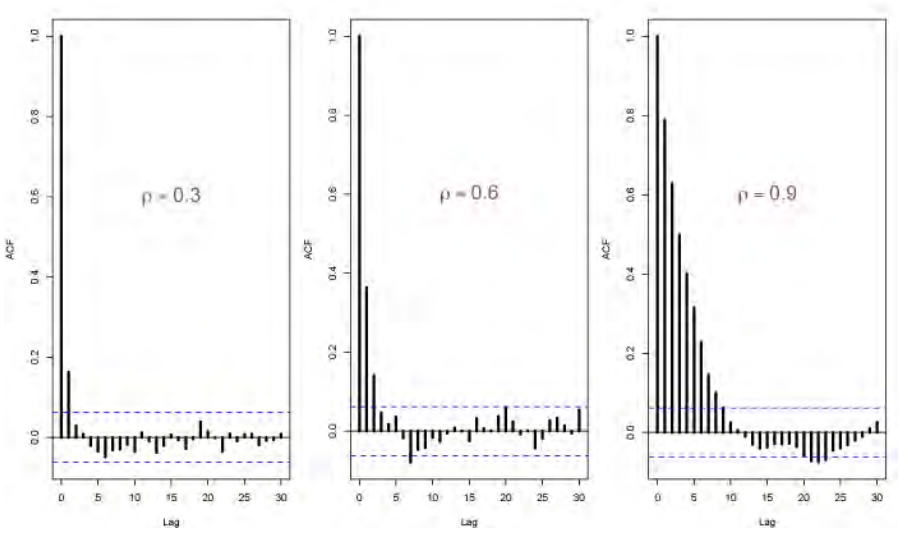


**Fig. 7.11.** Autocorrelations in one marginal of a bivariate normal generated from a Gibbs sampler for $\rho = 0.3$ (left), $\rho = 0.6$ (middle), and $\rho = 0.9$ (right).

b. Suppose now that $X$ and $Y$ are bivariate normal with mean 0, correlation $\rho$, and $\mathrm{var}(X) = \sigma_x^2$ and $\mathrm{var}(T) = \sigma_y^2$, which are not necessarily equal. Study the effect on autocorrelation of varying $\rho$, $\sigma_x^2$, and $\sigma_y^2$.
c. If $\sigma_x^2 \neq \sigma_y^2$, then $X + Y$ and $X - Y$ are no longer independent. Find a pair of random variables that are.

Convergence of both Gibbs sampling and Metropolis–Hastings algorithms may thus suffer from a poor choice of parameterization. As a result of this, the MCMC literature has considered changes in the parameterization of a model as a way to speed up convergence in a Gibbs sampler. It seems, however, that most efforts have concentrated on the improvement of specific models, resulting in a lack of general methodology for the choice of a "proper" parameterization. Nevertheless, the overall advice is to try to make the components "as independent as possible" and to use several parameterizations simultaneously to intermingle the conditionals.

**Example 7.14. (Continuation of Example 7.5)** A reparameterization of the one-way random effect of Example 7.5 is to introduce the overall mean at the observation level, as in

$$\begin{aligned} X_{ij} &\sim \mathcal{N}(\mu + \theta_i, \sigma^2), \quad i = 1, \ldots, k, \quad j = 1, \ldots, n_i, \\ \theta_i &\sim \mathcal{N}(0, \tau^2), \quad i = 1, \ldots, k, \\ \mu &\sim \mathcal{N}(\mu_0, \sigma_\mu^2). \end{aligned} \tag{7.10}$$

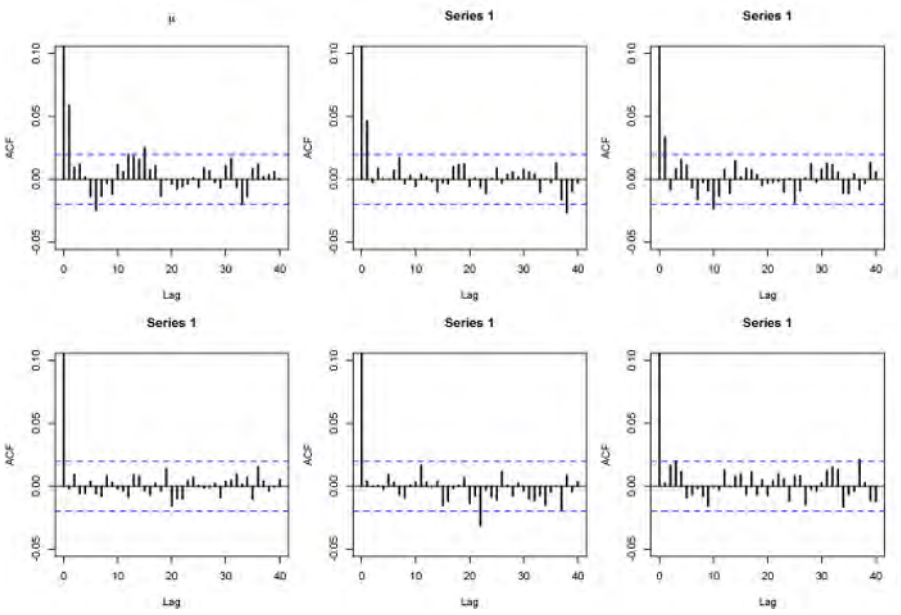


**Fig. 7.12.** Autocovariance plots for the Gibbs sampler associated with model (7.7) and the Gibbs sampler associated with its reparameterization (7.10). The top row gives the autocovariances for $\mu, \theta_1, \theta_2$ (left to right) for model (7.7), and the bottom row gives them for model (7.10).

While the hierarchy looks the same, the conditionals are different (Exercise 7.14) and the properties of the corresponding Gibbs sampler are as well. When applied to the `Energy` dataset, the new Gibbs sampler is not as good. For instance, Figure 7.12 shows the autocorrelations, which, to the eye, seem only slightly better for the first model. However, if we look at the covariance matrix of the subchain $(\mu^{(t)}, \theta_1^{(t)}, \theta_2^{(t)})$, its estimate is

$$\begin{pmatrix} 1.056 & -0.175 & -0.166 \\ -0.175 & 1.029 & 0.018 \\ -0.166 & 0.018 & 1.026 \end{pmatrix} \text{ and } \begin{pmatrix} 1.604 & 0.681 & 0.698 \\ 0.681 & 1.289 & 0.278 \\ 0.698 & 0.278 & 1.304 \end{pmatrix},$$

for model (7.7) and model (7.10), respectively, so the variances and covariances are larger for the reparameterized model. Thus, we clearly should use the parameterization of model (7.7). ◀

**Exercise 7.14** For the reparameterized model of (7.10):

a. Show that the full conditionals of $\theta_i$ and $\mu$ are

$$\theta_i \sim \mathcal{N}\left(B_1(\bar{X}_i - \mu), (\sigma^2/n_i)B_1\right), \quad B_1 = \frac{n_i\tau^2}{n_i\tau^2 + \sigma^2}, \quad i = 1, \ldots, k$$

$$\mu \sim \mathcal{N}\left((1 - B_2)\mu_0 + B_2(\bar{X} - \bar{\theta}), (\sigma^2/n)B_2\right), \quad B_2 = \frac{n\sigma_\mu^2}{n\sigma_\mu^2 + \sigma^2},$$

where $n = \sum_i n_i$ and $\bar{\theta} = \sum_i n_i\theta_i/n$.

b. Write a Gibbs sampler for this model, and compare the autocovariances with those of the Gibbs sampler based on model (7.7).

c. The covariance matrix of the parameter estimates is the inverse of the Fisher information matrix. Calculate this matrix for both parameterizations using the `R` functions `cor` and `solve`.

### 7.6.2 Rao–Blackwellization

We have already seen Rao–Blackwellization in Section 4.6, where conditioning on a subset of the simulated variables may produce considerable improvement upon the standard empirical estimator in terms of variance by a simple "recycling" of the rejected variables. However, as the Gibbs sampler accepts every simulated value, this type of recycling cannot apply. Nonetheless, Gelfand and Smith (1990) propose a type of conditioning that we will call *parametric Rao–Blackwellization* to differentiate it from the form studied in Section 4.6.

For $(X, Y) \sim f(x, y)$, *parametric Rao–Blackwellization* is based on the marginalization identity (iterated expectation)

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]].$$

Defining $\delta(Y) = \mathbb{E}[X|Y]$, we have $\mathbb{E}[\delta(Y)] = \mathbb{E}[X]$ and $\text{var}[\delta(Y)] \leq \text{var}(X)$, showing that $\delta(Y)$ is the better estimator (provided it can be computed).

**Example 7.15. (Continuation of Example 7.1)** In the case where the target is the bivariate normal distribution, the full conditionals are

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

and thus it follows that $\mathbb{E}[X|Y] = \rho Y$. Since $X$ and $Y$ have the same marginal distribution, the variance of the Rao–Blackwellized version is then obviously reduced by a factor $\rho^2$. ◀

Unfortunately, the variance reduction from using $\delta_Y$ does not hold in general, due to the correlation in the MCMC sample. However, Liu et al. (1994) have shown that, in particular, the improvement holds for any two-stage Gibbs sampler.

We now look at another example of Rao–Blackwellization in a missing-data Gibbs sampler for a common occurrence where possible gains occur.

**Example 7.16.** For 360 consecutive time units, consider recording the number of passages of individuals per unit time past some sensor. This can be, for instance, the number of cars observed at a crossroad. Hypothetical results are

| Number of passages | 0 | 1 | 2 | 3 | 4 or more |
|---|---|---|---|---|---|
| Number of observations | 139 | 128 | 55 | 25 | 13 |

The data involves a grouping of the observations with four passages or more. This can be addressed as a missing-data model, where we assume that the ungrouped observations are $X_i \sim \mathcal{P}(\lambda)$. The likelihood of the model is

$$\ell(\lambda|x_1,\ldots,x_5) \propto e^{-347\lambda}\lambda^{128+55\times 2+25\times 3}\left(1-e^{-\lambda}\sum_{i=0}^{3}\lambda^i/i!\right)^{13}$$

for $x_1 = 139,\ldots,x_5 = 13$. For $\pi(\lambda) = 1/\lambda$ and $\mathbf{z} = (z_1,\ldots,z_{13})$, the vector of the 13 units larger than 4, we can derive a completion Gibbs sampler from the full conditionals

$$Z_i^{(t)} \sim \mathcal{P}(\lambda^{(t-1)})\,\mathbb{I}_{y\geq 4}, \quad i = 1,\ldots,13,$$
$$\lambda^{(t)} \sim \mathcal{G}\left(313+\sum_{i=1}^{13} Z_i^{(t)},\ 360\right).$$

The Rao–Blackwellized estimate of $\lambda$ is then given by

$$\sum_{t=1}^{T}\mathbb{E}\left[\lambda\middle|, z_1^{(t)},\ldots,z_{13}^{(t)}\right] = \frac{1}{360T}\sum_{t=1}^{T}\left(313+\sum_{i=1}^{13} y_i^{(t)}\right),$$

and the evolution of this estimator, along with the empirical average, is shown in Figure 7.13. It exhibits a massive variance reduction. ◀

**Exercise 7.15** Referring to Example 7.16:

a. Verify the likelihood function and the Gibbs sampler.
b. Write `R` code to reproduce Figure 7.13.
c. The truncated Poisson variable can be generated using the `while` statement

```
> for (i in 1:13){while(y[i]<4) y[i]=rpois(1,lam[j-1])}
```

or directly with

```
> prob=dpois(c(4:top),lam[j-1])
> for (i in 1:13) z[i]=4+sum(prob<runif(1)*sum(prob))
```

Compare the efficiencies of these two algorithms. In theory, the value of `top` should be infinity. In practice, what value would you use?

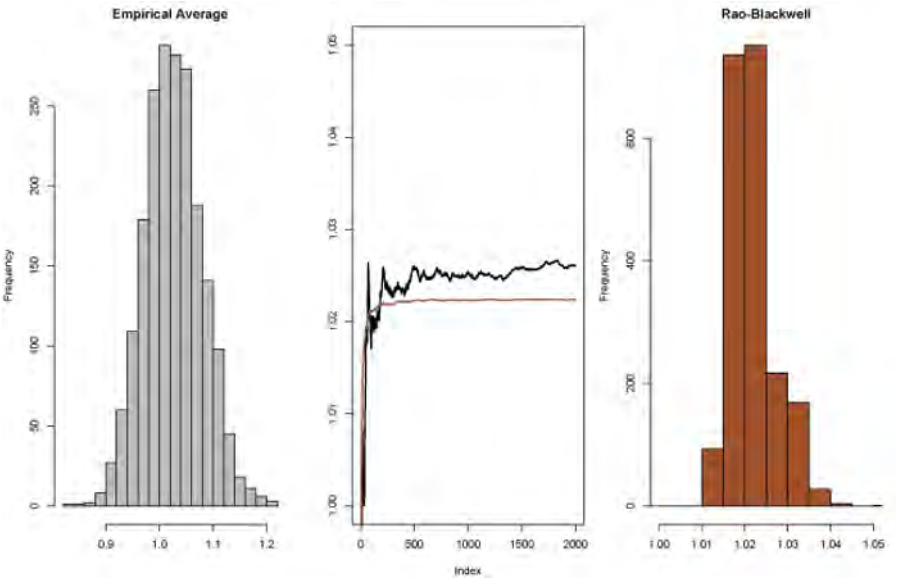


**Fig. 7.13.** For the counting data of Example 7.16, the histogram of $\lambda$ (left) and its conditional expectation $\mathbb{E}(\lambda|Z) = 313 + \sum_{i=1}^{13} z_i$ (right). Note the difference in scale of the histograms. The center panel shows the evolution of the cumulative averages of the empirical average (black) and the Rao–Blackwellization (grey).

There also exist non-parametric Rao–Blackwellized estimators in missing-variable settings. When considering an approximation to the marginal distribution $f_X$ associated with $f(x, y_1, \ldots, y_p)$, a Rao–Blackwellized estimator associated with the Gibbs chain $(x^{(t)}, \mathbf{y}^{(t)})_t$ is given by

$$\widehat{f}_X(x) = \frac{1}{T} \sum_{t=1}^{T} f(x|\mathbf{y}^{(t)}), \tag{7.11}$$

which converges at a parametric speed to $f_X(x)$. This estimator gives a smooth approximation to the marginal, which can be plotted on top of the marginal.

As studied in Exercises 3.15, 4.1, and 4.2, the approximation of the Bayes factors calls for specific solutions. Chib (1995) proposes an alternative approach based on a Rao–Blackwellization that is much more efficient when it can be implemented.

**Exercise 7.16** In a missing-variable setting where the sampling density can be written as

$$f(x|\theta) = \int_{\mathcal{Z}} g(x, z|\theta)\, \mathrm{d}z,$$

we assume the prior $\pi(\theta)$ is such that a two-stage Gibbs sampler based on the simulation of $g(z|x, \theta)$ and $\pi(\theta|x, z)$ can be implemented. Using a Bayes' Theorem

representation of the marginal density,

$$m(x) = \frac{f(x|\theta)\pi(\theta)}{\pi(\theta|x)}\,,$$

deduce a converging estimator of $m(x)$ based on the Rao–Blackwellized estimate of the posterior density $\pi(\theta|x)$ above. Apply to the settings of Examples 7.7 and 7.9.

### 7.6.3 Metropolis within Gibbs and hybrid strategies

A point worth emphasizing about the implementation of a Gibbs sampler is that it can easily be extended to settings where some of the full conditionals cannot be simulated by standard random generators. If, within a set of full conditionals $f_1, \ldots, f_p$, some density $f_i$ is unconventional, for example (5.15) in Example 5.17, this does not jeopardize the resulting Gibbs sampler in the sense that the following *Metropolis-within-Gibbs* strategy can be adopted: Instead of simulating

$$X_i^{(t+1)} \sim f_i(x_i|x_1^{(t+1)}, \ldots, x_{i-1}^{(t+1)}, x_{i+1}^{(t)}, \ldots, x_p^{(t)})\,,$$

you can run *one single step* of any MCMC scheme associated with the stationary distribution $f_i(x_i|x_1^{(t+1)}, \ldots, x_{i-1}^{(t+1)}, x_{i+1}^{(t)}, \ldots, x_p^{(t)})$. A simple solution is for instance to use a random walk Metropolis algorithm centered at $x_i^{(t)}$. While at first this sounds like a crude approximation, as the full conditional is not *exactly* simulated, the validity of the resulting algorithm is exactly the same as with the original Gibbs sampler since the joint distribution $f$ remains the stationary distribution of the corresponding Markov chain.

You may then wonder what the point is in using a Gibbs sampler if componentwise simulations have to be replaced with Metropolis–Hastings steps, as using a Metropolis–Hastings algorithm targeted at the joint distribution $f$ is more "natural". While there is nothing restraining you from using a joint Metropolis–Hastings algorithm, it is most often the case that designing such a Metropolis–Hastings algorithm on a large-dimensional target is challenging or even impossible. The fundamental gain in using a Gibbs-like structure is that it breaks down a complex model into a large number of smaller *and* simpler targets, where local Metropolis–Hastings algorithms can be designed at little expense.

**Example 7.17.** If we consider the target distribution (5.15), we mentioned in Example 5.17 that this is not a standard distribution. While Booth and Hobert (1999) designed a specific Accept-Reject algorithm to simulate from (5.15), a random walk proposal on each $u_i$, as in

```
> for (i in 1:n){
+  mu=u[i]
+  u[i]=factor*sigma[iter-1]*rnorm(1)+mu
+  if (log(runif(1))>gu(u[i],i,beta[iter-1],sigma[iter-1])-
+                      gu(mu,i,beta[iter-1],sigma[iter-1])){
+        u[i]=mu
+        }
+  }
```

produces a sample of $u_i$'s at iteration `iter` conditional on the current values of the parameters and the sample of $u_i$'s at iteration `iter-1`. In the overall Gibbs sampler, the parameters are then simulated by

```
> sigma=c(sigma,1/sqrt(2*rgamma(1,0.5*n)/sum(u^2)))
> tau=sigma[iter]/sqrt(sum(as.vector(x^2)*pro(beta[iter-1],u)))
> betaprop=beta[iter-1]+rnorm(1)*factor*tau
> if (log(runif(1))>likecomp(betaprop,sigma[iter],u)-
+                 likecomp(beta[iter-1],sigma[iter],u))
+     betarop=beta[iter-1]
> beta=c(beta,betaprop)
```

in a straightforward manner. (See Example 8.1 for the complete implementation.) The calibration term `factor` can further be tuned against the acceptance rate of Section 6.5, as described in Section 8.5. ◀

While remaining close to this idea of incorporating Metropolis–Hastings steps when direct simulation is not possible, we may also signal the possible extension to *hybrid strategies*.[4] The concept is once again based on the stationarity of the right target distribution, even though intuition may disagree. When given a (univariate or multivariate) target where several natural MCMC schemes are available, a hybrid algorithm merges those different schemes altogether. Schematically, if local or global (meaning componentwise or joint) MCMC update functions `mcmc.1(x,y)`, ..., `mcmc.q(x,y)` are available, the transition kernel defined by

```
mcmc(x,y)=function(x,y){
  switch(sample(1:p,1),
  mcmc.1(x,y)
     ...
  mcmc.p(x,y))
  }
```

remains a valid MCMC update function against the *same* target distribution. While this sounds like a ludicrous idea because poor schemes are mixed

[4] Hybrid strategies should not be confused with *hybrid Monte Carlo* (Neal, 1999), also called Hamiltonian MCMC, which is a form of Langevin implementation aimed at reducing the waste of simulation in random walk proposals.

with good ones, the blind mixing of all available strategies is nonetheless (a) valid from the perspective of producing the correct stationary distribution and (b) risk-free in the sense that if the list of functions contains a single well-performing algorithm, the hybrid version will perform at least as well, simply requiring a $p$-fold extension of the computing time. For instance, if several blocking or reparameterization strategies are simultaneously available, they can all be incorporated within the same algorithm. This solution could well appear as a waste of computing time, but our advice on this matter is that, unless some of the `mcmc.i` functions clearly do work, the time spent (wasted) running the hybrid solution is time saved on designing and selecting the more efficient `mcmc.i` functions. In other words, it is more efficient to let the computer sort among the available solutions than to run preliminary tests to sort those solutions "by hand".

### 7.6.4 Improper priors

This section discusses a particular danger resulting from careless use of the Gibbs sampler. We know that the Gibbs sampler is based on conditional distributions derived from the joint distribution. However, what is particularly insidious is that these conditional distributions may be well-defined and may be simulated from but may not correspond to any joint distribution!

This problem is not a *defect* of the Gibbs sampler, or even a simulation problem, but rather a problem of inadvertently using the Gibbs sampler in a situation for which the underlying assumptions are violated. It is nonetheless important to warn the user of MCMC algorithms against this danger because it corresponds to a situation often encountered in Bayesian noninformative (or *"default"*) models.

The construction of the Gibbs sampler directly from the conditional distributions is a strong incentive to bypass checking for the propriety of the posterior, especially in complex setups. But such checking is essential, as the following simple example shows.

**Example 7.18.** The following model was used by Casella and George (1992) to point out the difficulty of assessing the impropriety of a posterior distribution through the conditional distributions. The pair of conditional densities

$$X|y \sim \mathcal{E}xp(y)\,, \qquad Y|x \sim \mathcal{E}xp(x)\,, \tag{7.12}$$

are well-defined conditional distributions, but these conditional distributions do not correspond to any joint probability distribution. Figure 7.14 shows a histogram and cumulative average for a sample generated using the Gibbs sampler corresponding to those conditionals. The pictures are extremely curious and in fact are absolute rubbish! (This is not a recurrent Markov chain.) Indeed, the only function that could be the joint distribution is

$$f(x, y) \propto \exp(-xy),$$

which does not have a finite integral. ◀

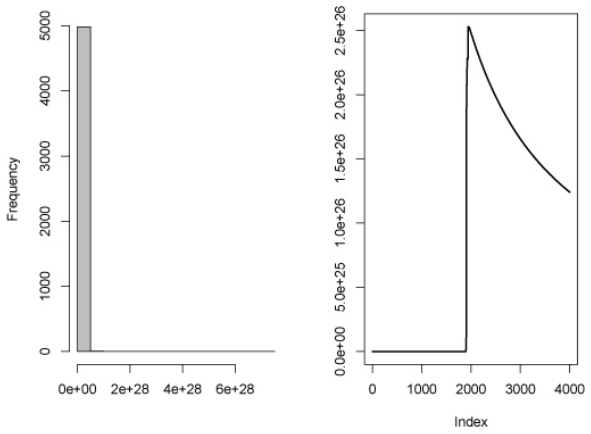


**Fig. 7.14.** Histogram and cumulative average of the $X$ variable from the Gibbs sampler of (7.12). Note the ranges on the graphs that signal convergence issues.

**Exercise 7.17** For the Gibbs sampler based on (7.12)

a. Write an R program to reproduce Figure 7.14.
b. The Hammersley–Clifford Theorem (Robert and Casella, 2004, Section 9.1.4) says that the joint density must satisfy

$$f(x,y) = f(y|x) \Big/ \int [f(y|x)/f(x|y)]\, \mathrm{d}y\,.$$

Show that applying this result to (7.12) leads to $f(x,y) \propto \exp(-xy)$.
c. Show that if the exponential distributions are restricted to $(0,B)$, $B < \infty$, the resulting figure is reasonable. Exhibit the stationary density of the Markov chain in this case. (*Hint:* Apply the Hammersley–Clifford Theorem.)

Given the results of Example 7.18, it may appear that a simple graphical monitoring is enough to exhibit deviant behavior of the Gibbs sampler. However, this is not the case in general and there are many examples, some of which are published (see Casella, 1996), where the output of the Gibbs sampler seemingly does not differ from a convergent Markov chain. Often, this phenomenon takes place when the divergence of the posterior density occurs "at 0"; that is, at a specific point whose immediate neighborhood is rarely visited by the chain, as in the following random effects example. The only way to make sure the Gibbs sampler you are using is valid is to check that the joint distribution has a finite integral.

**Example 7.19.** Consider a random effects model,

$$Y_{ij} = \beta + U_i + \varepsilon_{ij}, \qquad i = 1, \ldots, I,\ j = 1, \ldots, J,$$

where $U_i \sim \mathcal{N}(0, \sigma^2)$ and $\varepsilon_{ij} \sim \mathcal{N}(0, \tau^2)$. The Jeffreys (improper) prior for the parameters $\beta$, $\sigma$, and $\tau$ is

$$\pi(\beta, \sigma^2, \tau^2) = \frac{1}{\sigma^2 \tau^2}\ .$$

The conditional distributions

$$\begin{aligned}
U_i|y, \beta, \sigma^2, \tau^2 &\sim \mathcal{N}\left(\frac{J(\bar{y}_i - \beta)}{J + \tau^2\sigma^{-2}}, (J\tau^{-2} + \sigma^{-2})^{-1}\right), \\
\beta|u, y, \sigma^2, \tau^2 &\sim \mathcal{N}(\bar{y} - \bar{u}, \tau^2/JI), \\
\sigma^2|u, \beta, y, \tau^2 &\sim \mathcal{IG}\left(I/2, (1/2)\sum_i u_i^2\right), \\
\tau^2|u, \beta, y, \sigma^2 &\sim \mathcal{IG}\left(IJ/2, (1/2)\sum_{i,j} (y_{ij} - u_i - \beta)^2\right),
\end{aligned}$$

are well-defined, and a Gibbs sampler can be easily implemented in this setting. However, there is no proper joint distribution that corresponds to these conditionals! And, in many instances, as you may check for yourself, this is impossible to detect by monitoring the output. ◀

**Exercise 7.18** In the setting of Example 7.19:

a. Generate data according to the model and run a corresponding Gibbs sampler on the parameters of the model. Monitor histograms and cumulative averages. Can you detect the fact that there is no proper joint distribution?
b. The variation on the model (7.6) given in Exercise 7.6, where $\mu$ is given a flat prior, is a Gibbs sampler with improper priors. Since there is no guarantee that the posterior distribution is proper, check to see if it is in fact proper.

ϟ If improper priors are used in a Gibbs sampler, the posterior must *always* be checked for propriety. However, it is often the case that improper priors on variances cause more trouble than those on means.

## 7.7 Additional exercises

**Exercise 7.19** The gamma distribution with parameters $a$ and $b$, $\mathcal{G}(a, b)$, has density $b^a x^{a-1} e^{-bx}/\Gamma(a)$. Show that if $X \sim \mathcal{IG}(a, b)$, then $1/X \sim \mathcal{G}(a, b)$. (This means that

generating from a gamma distribution is equivalent to generating from an inverted gamma distribution.)

**Exercise 7.20** From the hierarchy (7.6), show that the joint distribution can be obtained by multiplying the densities together. Then, using the strategy of Exercise 7.4, verify that the full conditionals are given by (7.7).

**Exercise 7.21** A truncated normal generator is based on the `R` function

```
rtnorm=function(n=1,mu=0,lo=-Inf,up=Inf){
  qnorm(runif(n,min=pnorm(lo,mean=mu,sd=sigma),
    max=pnorm(up,mean=mu,sd=sigma)),
    mean=mu,sd=sigma)}
```

where `mu` and `sigma` are the mean and standard deviation of the normal, `lo` is the lower truncation point, `up` is the upper truncation point, and `n` is the number of random variables desired. For $Z \sim \mathcal{N}(0,1)$ with truncation $(i)$ $-1 < Z < 1$, $(ii)$ $Z < 1$, and $(iii)$ $Z > 3$, generate $1000$ random variables and compare the histograms with the density functions.

**Exercise 7.22** Referring to Exercise 7.5:

a. Calculate the third and fourth moments of the density in question a of that exercise.
b. If $\mathbf{X} \sim \mathcal{N}_p(0, \Sigma)$, show that the density of $X_1|x_{(-1)}$ is

$$\mathcal{N}_p(\Sigma_{12}\Sigma_{22}^{-1}x_{(-1)}, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}'),$$

where the covariance matrix is partitioned in the obvious way. Use this formula to verify (7.5).
c. The matrix $(1-\rho)I + \rho J$ is only positive definite if $\rho > -1/(p-2)$. Verify this result.

**Exercise 7.23** Referring to model (7.8), the (uncompleted) posterior distribution is available as

$$\begin{aligned}\pi(\theta_1, \theta_2|x) \propto\ & (a_1\theta_2 + b_1)^{x_1}(a_2\theta_2 + b_2)^{x_2}(a_3\theta_1 + b_3)^{x_3}(a_4\theta_1 + b_4)^{x_4} \\ & \times (1 - \theta_1 - \theta_2)^{x_5+\alpha_3-1}\theta_2^{\alpha_1-1}\theta_1^{\alpha_2-1}.\end{aligned}$$

a. Show that the marginal distributions $\pi(\theta_1|x)$ and $\pi(\theta_2|x)$ can be explicitly computed as polynomials when the $\alpha_i$'s are integers.
b. Give the marginal posterior distribution of $\xi = \theta_2/(1-\theta_1-\theta_2)$. (*Note:* See Robert, 1995a, for a solution.)
c. Evaluate the Gibbs sampler based on (7.9) by comparing approximate moments of $\theta_1$, $\theta_1$, and $\xi$ with their exact counterparts derived from the explicit marginal.

**Exercise 7.24** The alternate parameterization of model (7.6) produced in Example 7.14 modifies the relations between the variables. Show that $\theta_i$ and $\mu$ are a priori independent for this parameterization and that this is not the case in model (7.6).

**Exercise 7.25** Rao–Blackwellization can be applied to most of the Gibbs samplers in this chapter. For each of the following examples, verify the conditional expectations provided there and compare via an `R` experiment the empirical average with the Rao–Blackwellization.

a. Example 7.2: $\mathbb{E}[\theta|x] = x + a/(n + a + b)$.
b. Equation (7.4): $\mathbb{E}[\theta|\mathbf{x}, \sigma^2] = \frac{\sigma^2}{\sigma^2+n\tau^2}\ \theta_0 + \frac{n\tau^2}{\sigma^2+n\tau^2}\ \bar{x}$.
c. Equation (7.7): $\mathbb{E}[\theta_i|\bar{X}_i, \sigma^2] = \frac{\sigma^2}{\sigma^2+n_i\tau^2}\mu + \frac{n_i\tau^2}{\sigma^2+n_i\tau^2}\bar{X}_i$.
d. Example 7.6: $\mathbb{E}[\theta|x, z] = \frac{m\bar{x}+(n-m)\bar{z}}{n}$.
e. Example 7.12: $\mathbb{E}[\lambda_i|\beta, t_i, x_i] = (x_i + \alpha)/(t_i + \beta)$.