4

Modern Model Estimation Part 1: Gibbs Sampling

The estimation of a Bayesian model is the most difficult part of undertaking a Bayesian analysis. Given that researchers may use different priors for any particular model, estimation must be tailored to the specific model under consideration. Classical analyses, on the other hand, often involve the use of standard likelihood functions, and hence, once an estimation routine is developed, it can be used again and again.

The trade-off for the additional work required for a Bayesian analysis is that (1) a more appropriate model for the data can be constructed than extant software may allow, (2) more measures of model fit and outlier/influential case diagnostics can be produced, and (3) more information is generally available to summarize knowledge about model parameters than a classical analysis based on maximum likelihood (ML) estimation provides. Along these same lines, additional measures may be constructed to test hypotheses concerning parameters not directly estimated in the model.

In this chapter, I first discuss the goal of model estimation in the Bayesian paradigm and contrast it with that of maximum likelihood estimation. Then, I discuss modern simulation/sampling methods used by Bayesian statisticians to perform analyses, including Gibbs sampling. In the next chapter, I discuss the Metropolis-Hastings algorithm as an alternative to Gibbs sampling.

4.1 What Bayesians want and why

As the discussion of ML estimation in Chapter 2 showed, the ML approach finds the parameter values that maximize the likelihood function for the observed data and then produces point estimates of the standard errors of these estimates. A typical classical statistical test is then conducted by subtracting a hypothesized value for the parameter from the ML estimate and dividing the result by the estimated standard error. This process yields a standardized estimate (under the hypothesized value). The Central Limit Theorem states that the sampling distribution for a sample statistic/parameter estimate is

asymptotically normal, and so we can use the z (or t) distribution to evaluate the probability of observing the sample statistic we observed under the assumption that the hypothesized value for it were true. If observing the sample statistic we did would be an extremely rare event under the hypothesized value, we reject the hypothesized value.

In contrast to the use of a single point estimate for a parameter and its standard error and reliance on the Central Limit Theorem, a Bayesian analysis derives the posterior distribution for a parameter and then seeks to summarize the entire distribution. As we discussed in Chapter 2, many of the quantities that may be of interest in summarizing knowledge about a distribution are integrals of it, like the mean, median, variance, and various quantiles. Obtaining such integrals, therefore, is a key focus of Bayesian summarization and inference.

The benefits of using the entire posterior distribution, rather than point estimates of the mode of the likelihood function and standard errors, are several. First, if we can summarize the entire posterior distribution for a parameter, there is no need to rely on asymptotic arguments about the normality of the distribution: It can be directly assessed. Second, as stated above, having the entire posterior distribution for a parameter available allows for a considerable number of additional tests and summaries that cannot be performed under a classical likelihood-based approach. Third, as discussed in subsequent chapters, distributions for the parameters in the model can be easily transformed into distributions of quantities that may be of interest but may not be directly estimated as part of the original model. For example, in Chapter 10, I show how distributions for hazard model parameters estimated via Markov chain Monte Carlo (MCMC) methods can be transformed into distributions of life table quantities like healthy life expectancy. Distributions of this quantity cannot be directly estimated from data but instead can be computed as a function of parameters from a hazard model. A likelihood approach that produces only point estimates of the parameters and their associated standard errors cannot accomplish this.

Given the benefits of a Bayesian approach to inference, the key question then is: How difficult is it to integrate a posterior distribution to produce summaries of parameters?

4.2 The logic of sampling from posterior densities

For some distributions, integrals for summarizing posterior distributions have closed-form solutions and are known, or they can be easily computed using numerical methods. For example, in the previous chapter, we determined the expected proportion of—and a plausible range for—votes for Kerry in the 2004 presidential election in Ohio, as well as the probability that Kerry would win Ohio, using known information about integrals of the beta distribution. We also computed several summaries using a normal approximation to the

posterior density, and of course, integrals of the normal distribution are well-known.

For many distributions, especially multivariate ones, however, integrals may not be easy to compute. For example, if we had a beta prior distribution on the variance of a normal distribution, the posterior distribution for the variance would not have a known form. In order to remedy this problem, Bayesians often work with conjugate priors, as we discussed in the previous chapter. However, sometimes conjugate priors are unrealistic, or a model may involve distributions that simply are not amenable to simple computation of quantiles and other quantities. In those cases, there are essentially two basic approaches to computing integrals: approximation methods and sampling methods.

Before modern sampling methods (e.g., MCMC) were available or computationally feasible, Bayesians used a variety of approximation methods to perform integrations necessary to summarize posterior densities. Using these methods often required extensive knowledge of advanced numerical methods that social scientists generally do not possess, limiting the usefulness of a Bayesian approach. For example, quadrature methods—which involve evaluating weighted points on a multidimensional grid—were often used. As another example, Bayesians often generated Taylor series expansions around the mode of the log-posterior distribution, and then used normal approximations to the posterior for which integrals are known. For multimodal distributions, Bayesians would often use approximations based on mixtures of normals. All of these approaches were methods of *approximation* and, hence, formed a foundation for criticizing Bayesian analysis. Of course, it is true that a Bayesian Central Limit Theorem shows that *asymptotically* most posterior distributions are normal (see Gelman et al. 1995 for an in-depth discussion of asymptotic normal theory in a Bayesian setting), but reliance on this theorem undermines a key benefit of having a complete posterior distribution: the lack of need to—and, in small samples, the inability to—rely on asymptotic arguments. I do not focus on these methods in this book.

Sampling methods constitute an alternative to approximation methods. The logic of sampling is that we can generate (simulate) a sample of size n from the distribution of interest and then use discrete formulas applied to these samples to approximate the integrals of interest. Under a sampling approach, we can estimate a mean by:

$$\int x f(x) dx \approx \frac{1}{n} \sum x$$

and the variance by:

$$\int (x - \mu)^2 f(x) dx \approx \frac{1}{n} \sum (x - \mu)^2.$$

Various quantiles can be computed empirically by noting the value of x for which $Q\%$ of the sampled values fall below it.

Thus, modern Bayesian inference typically involves (1) establishing a model and obtaining a posterior distribution for the parameter(s) of interest, (2) generating samples from the posterior distribution, and (3) using discrete formulas applied to the samples from the posterior distribution to summarize our knowledge of the parameters. These summaries are not limited to a single quantity but instead are virtually limitless. Any summary statistic that we commonly compute to describe a sample of data can also be computed for a sample from a posterior distribution and can then be used to describe it!

Consider, for example, the voting example from the previous chapter in which we specified a beta prior distribution for K, coupled with a binomial likelihood for the most recent polling data. In that example, the posterior density for K was a beta density with parameters $\alpha = 1498$ and $\beta = 1519$. Given that the beta density is a known density, we computed the posterior mean as $1498/(1498 + 1519) = .497$, and the probability that $K > .5$ as .351. However, assume these integrals could not be computed analytically. In that case, we could simulate several thousand draws from this particular beta density (using `x=rbeta(5000,1498,1519)`in R, with the first argument being the desired number of samples), and we could then compute the mean, median, and other desired quantities from this sample. I performed this simulation and obtained a mean of .496 for the 5,000 samples (obtained by typing `mean(x)` in R) and a probability of .351 that Kerry would win (obtained by typing `sum(x>.5)/5000)`.

Notice that the mean obtained analytically (via integration of the posterior density) and the mean obtained via sampling are identical to almost three decimal places, as are the estimated probabilities that Kerry would win. The reason that these estimates are close is that sampling methods, in the limit, are not approximations; instead, they provide exact summaries equivalent to those obtained via integration. A sample of 5,000 draws from this beta distribution is more than sufficient to accurately summarize the density. As a demonstration, Figure 4.1 shows the convergence of the sample-estimated mean for this particular beta distribution as the sample size increases from 1 to 100,000. At samples of size $n = 5,000$, the confidence band around the mean is only approximately .0005 units wide. In other words, our error in using simulation rather than analytic integration is extremely small. As the sample size increases, we can see that the simulation error diminishes even further.

4.3 Two basic sampling methods

In the example shown above, it was easy to obtain samples from the desired beta density using a simple command in R. For many distributions, there are effective routines in existence for simulating from them (some of which ultimately rely on the inversion method discussed below). For other distributions, there may not be an extant routine, and hence, a statistician may need

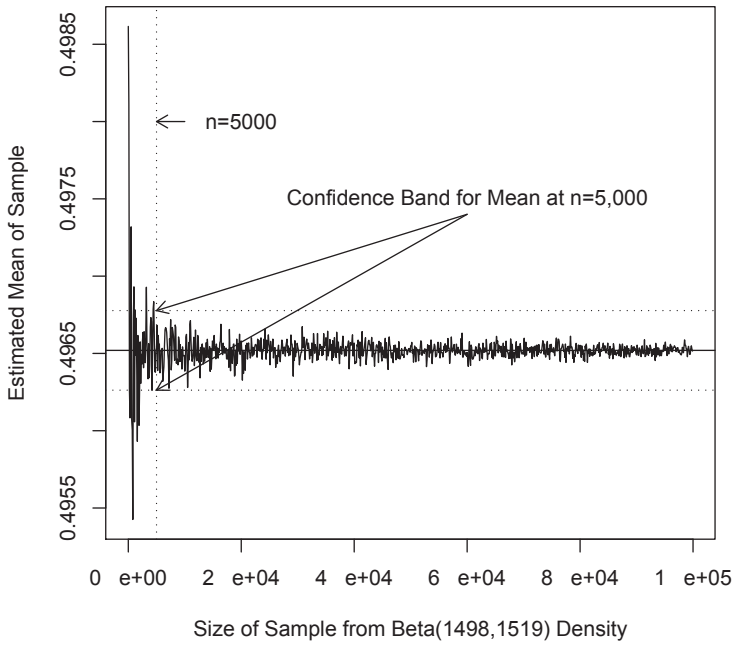

Fig. 4.1. Convergence of sample means on the true beta distribution mean across samples sizes: Vertical line shows sample size of 5,000; dashed horizontal lines show approximate confidence band of sample estimates for samples of size $n = 5,000$; and solid horizontal line shows the true mean.

to create one. Indeed, this is the entire reason for MCMC methods, as we will discuss: Integration of posterior densities is often impossible, and there may not be extant routines for sampling from them either, especially when they are high-dimensional. I first discuss two sampling methods, each of which is important for a basic understanding of MCMC methods. These methods, as well as several others, are described in greater depth in Gilks (1996). For a more detailed exposition on simulation methods, see Ripley (1987).

4.3.1 The inversion method of sampling

For drawing a sample from a univariate distribution $f(x)$, we can often use the inversion method. The inversion method is quite simple and follows two steps:

1. Draw a uniform random number u between 0 and 1 (a $U(0,1)$ random variable).

2. Then $z = F^{-1}(u)$ is a draw from $f(x)$.

In step 1, we draw a $U(0,1)$ random variable. This draw represents the area under the curve up to the value of our desired random draw from the distribution of interest. Thus, we simply need to find z such that:

$$u = \int_L^z f(x)dx,$$

where L is the lower limit of the density f. Put another way, $u = F(z)$. So, phrased in terms of z:

$$z = F^{-1}(u).$$

To provide a concrete example, take the linear density function from Chapter 2: $f(x) = (1/40)(2x+3)$ (with $0 < x < 5$). As far as I know, no routines are readily available that allow sampling from this density, and so, if one needed draws from this density, one would need to develop one. In order to generate a draw from this distribution using the inversion method, we first need to draw $u \sim U(0,1)$ and then compute z that satisfies

$$u = \int_0^z \frac{1}{40}(2x+3)dx.$$

We can solve this equation for z as follows. First, evaluate the integral:

$$40u = \left. x^2 + 3x \right|_0^z = z^2 + 3z.$$

Second, complete the square in z:

$$40u + \frac{9}{4} = z^2 + 3z + \frac{9}{4} = \left(z + \frac{3}{2}\right)^2.$$

Third, take the square root of both sides and rearrange to find z:

$$z = \frac{-3 \pm \sqrt{160u + 9}}{2}.$$

This result reveals two solutions for z; however, given that z must be between 0 and 5, only the positive root is relevant. If we substitute 0 and 1—the minimum and maximum values for u—we find that the range of z is $[0, 5]$ as it should be.

Figure 4.2 displays the results of an algorithm simulating 1,000 random draws from this density using the inversion method. The figures on the left-hand side show the sequence of draws from the $U(0,1)$ density, which are then inverted to produce the sequence of draws from the density of interest. The right-hand side of the figure shows the simulated and theoretical density functions. Notice how the samples from both densities closely follow, but do not exactly match, the theoretical densities. This error is sampling error, which diminishes as the simulation sample size increases.

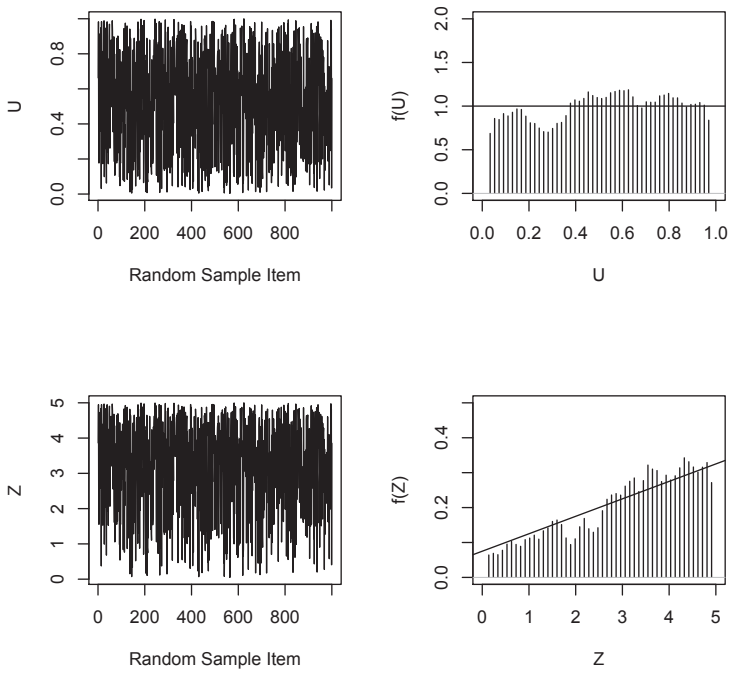

Fig. 4.2. Example of the inversion method: Left-hand figures show the sequence of draws from the $U(0,1)$ density (upper left) and the sequence of draws from the density $f(x) = (1/40)(2x+3)$ density (lower left); and the right-hand figures show these draws in histogram format, with true density functions superimposed.

The following R program was used to generate these draws. The first line simulates 1,000 random draws from the $U(0,1)$ distribution; the second line generates the vector z as the inverse of u :

```
#R program for inversion method of sampling
u=runif(1000,min=0,max=1)
z=(1/2) * (-3 + sqrt(160*u +9))
```

Although the inversion method is very efficient and easy to implement, two key limitations reduce its usability as a general method for drawing samples from posterior densities. First, if the inverse function is impossible to derive analytically, obviously the method cannot be used. For example, the normal integral cannot be directly solved, and hence, the inversion method cannot be used to simulate from the normal distribution.[1] To some extent, this problem

[1] Of course, we do have efficient algorithms for computing this integral, but the integral cannot be solved analytically.

begs the question: If we can integrate the density as required by the inversion method, then why bother with simulation? This question will be addressed shortly, but the short answer is that we may not be able to perform integration on a multivariate density, but we can often break a multivariate density into univariate ones for which inversion may work.

The second problem with the inversion method is that the method will not work with multivariate distributions, because the inverse is generally not unique beyond one dimension. For example, consider the bivariate planar density function discussed in Chapter 2:

$$f(x, y) = \frac{1}{28}(2x + 3y + 2),$$

with $0 < x, y < 2$. If we draw $u \sim U(0, 1)$ and attempt to solve the double integral for x and y, we get:

$$28u = yx^2 + \frac{3xy^2}{2} + 2xy,$$

which, of course, has infinitely many solutions (one equation with two unknowns). Thinking ahead, we could select a value for one variable and then use the inversion method to draw from the conditional distribution of the other variable. This process would reduce the problem to one of sampling from univariate conditional distributions, which is the basic idea of Gibbs sampling, as I discuss shortly.

4.3.2 The rejection method of sampling

When $F^{-1}(u)$ cannot be computed, other methods of sampling exist. A very important one is rejection sampling. In rejection sampling, sampling from a distribution $f(x)$ for x involves three basic steps:

1. Sample a value z from a distribution $g(x)$ from which sampling is easy and for which values of $m \times g(x)$ are greater than $f(x)$ at all points (m is a constant).
2. Compute the ratio $R = \frac{f(z)}{m \times g(z)}$.
3. Sample $u \sim U(0, 1)$. If $R > u$, then accept z as a draw from $f(x)$. Otherwise, return to step 1.

In this algorithm, $m \times g(x)$ is called an "envelope function," because of the requirement that the density function $g(x)$ multiplied by some constant m be greater than the density function value for the distribution of interest $[f(x)]$ at the same point for all points. In other words, $m \times g(x)$ *envelops* $f(x)$. In step 1, we sample a point z from the pdf $g(x)$.

In step 2, we compute the ratio of the envelope function $[m \times g(x)]$ evaluated at z to the density function of interest $[f(x)]$ evaluated at the same point.

Finally, in step 3, we draw a $U(0,1)$ random variable u and compare it with R. If $R > u$, then we treat the draw as a draw from $f(x)$. If not, we reject z as coming from $f(x)$, and we repeat the process until we obtain a satisfactory draw.

This routine is easy to implement, but it is not immediately apparent why it works. Let's again examine the density discussed in the previous section and consider an envelope function that is a uniform density on the $[0,5]$ interval multiplied by a constant of 2. I choose this constant because the height of the $U(0,5)$ density is .2, whereas the maximum height of the density $f(x) = (1/40)(2x+3)$ is .325. Multiplying the $U(0,5)$ density by two increases the height of this density to .4, which is well above the maximum for $f(x)$ and therefore makes $m \times g(x)$ a true envelope function. Figure 4.3 shows the density and envelope functions and graphically depicts the process of rejection sampling.

In the first step, when we are sampling from the envelope function, we are choosing a location on the x axis in the graph (see top graph in Figure 4.3). The process of constructing the ratio R and comparing it with a uniform deviate is essentially a process of locating a point in the y direction once the x coordinate is chosen and then deciding whether it is *under* the density of interest. This becomes more apparent if we rearrange the ratio and the inequality with u:

$$f(z) \underbrace{<=>}_{?} m \times g(z) \times u.$$

$m \times g(z) \times u$ provides us a point in the y dimension that falls somewhere between 0 and $m \times g(z)$. This can be easily seen by noting that $m \times g(z) \times u$ is really simply providing a random draw from the $U(0, g(z))$ distribution: The value of this computation when $u = 0$ is 0; its value when $u = 1$ is $m \times g(z)$ (see middle graph in Figure 4.3). In the last step, in which we decide whether to accept z as a draw from $f(x)$, we are simply determining whether the y coordinate falls below the $f(x)$ curve (see bottom graph in Figure 4.3). Another way to think about this process is that the ratio tells us the proportion of times we will accept a draw at a given value of x as coming from the density of interest.

The following R program simulates 1,000 draws from the density $f(x) = (1/40)(2x+3)$ using rejection sampling. The routine also keeps a count of how many total draws from $g(x)$ must be made in order to obtain 1,000 draws from $f(x)$.

```
#R program for rejection method of sampling
count=0; k=1; f=matrix(NA,1000)
while(k<1001)
 {
  z=runif(1,min=0,max=5)
  r=((1/40)*(2*z+3))/(2*.2)
```

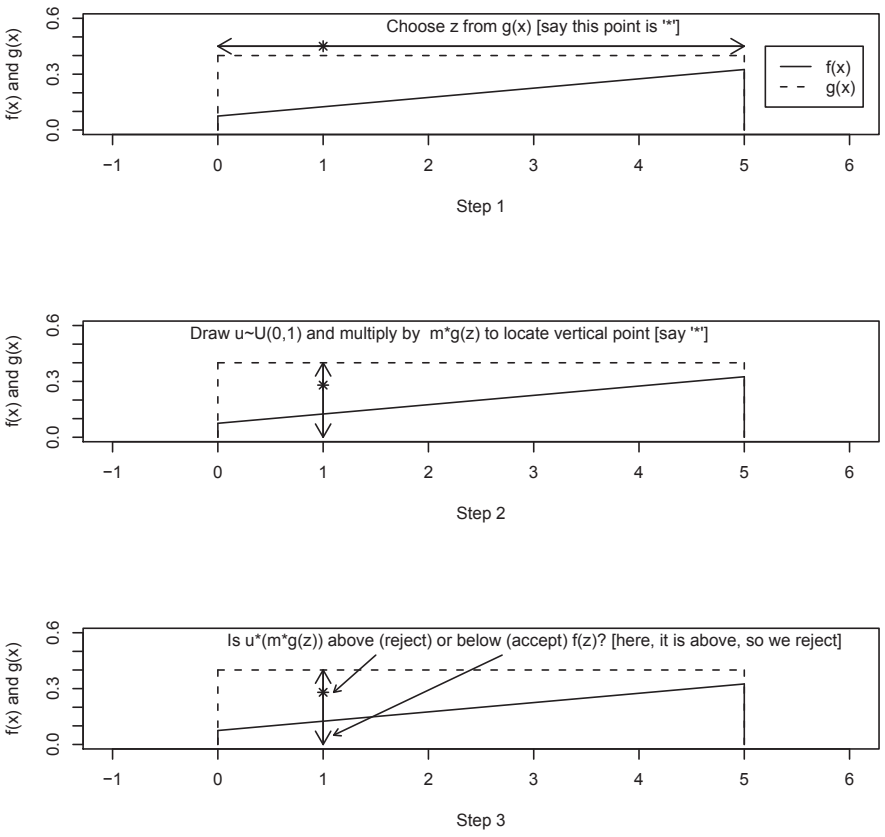

Fig. 4.3. The three-step process of rejection sampling.

```
 if(r>runif(1,min=0,max=1))
  {f[k]=z; k=k+1}
 count=count+1
}
```

Figure 4.4 shows the results of a run of this algorithm. The histogram of the sample of 1,000 draws very closely matches the density of interest.

Rejection sampling is a powerful method of sampling from densities for which inversion sampling does not work. It can be used to sample from *any* density, and it can be used to sample from multivariate densities. In the multivariate case, we first choose an X—now a random vector, rather than a single point—from a multivariate enveloping function, and then we proceed just as before.

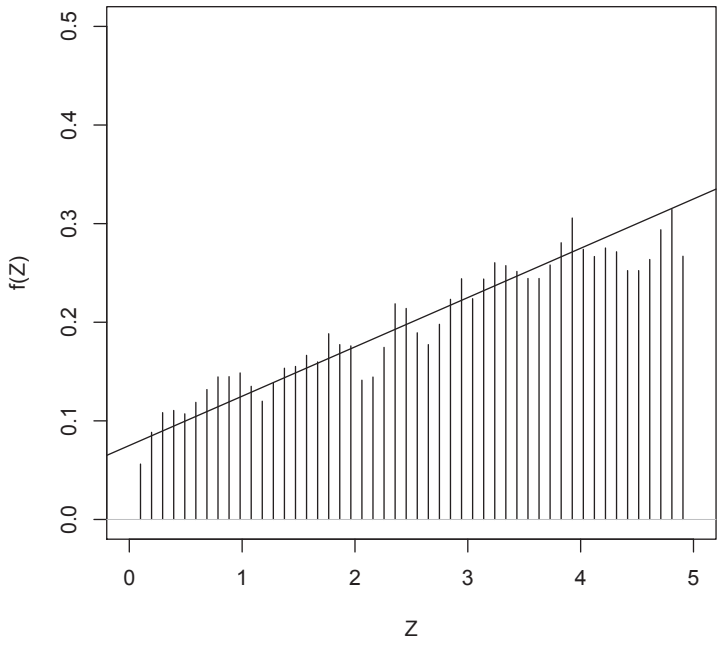

Fig. 4.4. Sample of 1,000 draws from density using rejection sampling with theoretical density superimposed.

Rejection sampling does have some limitations. First, finding an enveloping function $m \times g(x)$ may not be an easy task. For example, it may be difficult to find an envelope with values that are greater at all points of support for the density of interest. Consider trying to use a uniform density as an envelope for sampling from a normal density. The domain of x for the normal density runs from $-\infty$ to $+\infty$, but there is no corresponding uniform density. In the limit, a $U(-\infty, +\infty)$ density would have an infinitely low height, which would make $g(x)$ fall below $f(x)$ in the center of the distribution, regardless of the constant multiple m chosen. Second, the algorithm may not be very efficient. If the enveloping function is considerably higher than $f(x)$ at all points, the algorithm will reject most attempted draws, which implies that an incredible number of draws may need to be made before finding a single value from $f(x)$. In theory, the efficiency of a rejection sampling routine is calculable before implementing it. In the case above, the total area under the enveloping curve is 2 ($5 \times .4$), but the total area under the density of interest is 1 (by definition of a density function). Thus, the algorithm used should accept about 50% of the draws from $g(x)$. In fact, in the case shown and discussed above, it took 2,021

attempts to obtain 1,000 draws from $f(x)$, which is a rejection rate of 50.5%. These two limitations make rejection sampling, although possible, increasingly difficult as the dimensionality increases in multivariate distributions.

4.4 Introduction to MCMC sampling

The limitations of inversion and rejection sampling make the prospects of using these simple methods daunting in complex statistical analyses involving high-dimensional distributions. Although rejection sampling approaches can be refined to be more efficient, they are still not very useful in and of themselves in real-world statistical modeling. Fortunately, over the last few decades, MCMC methods have been developed that facilitate sampling from complex distributions. Furthermore, aside from allowing sampling from complex distributions, these methods provide several additional benefits, as we will be discussing in the remaining chapters.

MCMC sampling provides a method to sample from multivariate densities that are not easy to sample from, often by breaking these densities down into more manageable univariate or multivariate densities. The basic MCMC approach provides a prescription for (1) sampling from one or more dimensions of a posterior distribution and (2) moving throughout the entire support of a posterior distribution. In fact, the name "Markov chain Monte Carlo" implies this process. The "Monte Carlo" portion refers to the random simulation process. The "Markov chain" portion refers to the process of sampling a new value from the posterior distribution, given the previous value: This iterative process produces a Markov chain of values that constitute a sample of draws from the posterior.

4.4.1 Generic Gibbs sampling

The Gibbs sampler is the most basic MCMC method used in Bayesian statistics. Although Gibbs sampling was developed and used in physics prior to 1990, its widespread use in Bayesian statistics originated in 1990 with its introduction by Gelfand and Smith (1990). As will be discussed more in the next chapter, the Gibbs sampler is a special case of the more general Metropolis-Hastings algorithm that is useful when (1) sampling from a multivariate posterior is not feasible, but (2) sampling from the conditional distributions for each parameter (or blocks of them) is feasible. A generic Gibbs sampler follows the following iterative process (j indexes the iteration count):

0. Assign a vector of starting values, S, to the parameter vector: $\Theta^{j=0} = S$.
1. Set $j = j + 1$.
2. Sample $(\theta_1^j \mid \theta_2^{j-1}, \theta_3^{j-1} \ldots \theta_k^{j-1})$.
3. Sample $(\theta_2^j \mid \theta_1^j, \theta_3^{j-1} \ldots \theta_k^{j-1})$.
⋮ ⋮
k. Sample $(\theta_k^j \mid \theta_1^j, \theta_2^j, \ldots, \theta_{k-1}^j)$.
k+1. Return to step 1.

In other words, Gibbs sampling involves ordering the parameters and sampling from the conditional distribution for each parameter given the current value of all the other parameters and repeatedly cycling through this updating process. Each "loop" through these steps is called an "iteration" of the Gibbs sampler, and when a new sampled value of a parameter is obtained, it is called an "updated" value.

For Gibbs sampling, the full conditional density for a parameter needs only to be known up to a normalizing constant. As we discussed in Chapters 2 and 3, this implies that we can use the joint density with the other parameters set at their current values. This fact makes Gibbs sampling relatively simple for most problems in which the joint density reduces to known forms for each parameter once all other parameters are treated as fixed.

4.4.2 Gibbs sampling example using the inversion method

Here, I provide a simple example of Gibbs sampling based on the bivariate plane distribution developed in Chapter 2 $f(x, y) = (1/28)(2x + 3y + 2)$. The conditional distribution for x was:

$$f(x \mid y) = \frac{f(x, y)}{f(y)} = \frac{2x + 3y + 2}{6y + 8},$$

and the conditional distribution for y was:

$$f(y \mid x) = \frac{f(x, y)}{f(x)} = \frac{2x + 3y + 2}{4x + 10}.$$

Thus, a Gibbs sampler for sampling x and y in this problem would follow these steps:

1. Set $j = 0$ and establish starting values. Here, let's set $x^{j=0} = -5$ and $y^{j=0} = -5$.
2. Sample x^{j+1} from $f(x \mid y = y^j)$.
3. Sample y^{j+1} from $f(y \mid x = x^{j+1})$.
4. Increment $j = j + 1$ and return to step 2 until $j = 2000$.

How do we sample from these conditional distributions? We know what they are, but they certainly are not standard distributions. Since they are not standard distributions, but since these conditionals are univariate and $F^{-1}()$ can be calculated for each one, we can use an inversion subroutine to sample from each conditional density. How do we find the inverses in this bivariate density? Recall that inversion sampling requires first drawing a $u \sim U(0,1)$ random variable and then inverting this draw using F^{-1}. Thus, to find the inverse of the conditional density for $y|x$, we need to solve:

$$u = \int_0^z \frac{2x + 3y + 2}{4x + 10}$$

for z. Given that this is the conditional density for y, x is fixed and can be treated as a constant, and we obtain:

$$u(4x + 10) = (2x + 2)y + (3/2)y^2\big|_0^z.$$

Thus:

$$u(4x + 10) = (2x + 2)z + (3/2)z^2.$$

After multiplying through by $(2/3)$ and rearranging terms, we get:

$$(2/3)u(4x + 10) = z^2 + (2/3)(2x + 2)z.$$

We can then complete the square in z and solve for z to obtain:

$$z = \sqrt{(2/3)u(4x + 10) + ((1/3)(2x + 2))^2} - (1/3)(2x + 2).$$

Given a current value for x and a random draw u, z is a random draw from the conditional density for $y|x$. A similar process can be undertaken to find the inverse for $x|y$ (see Exercises).

Below is an R program that implements the Gibbs sampling:

```
#R program for Gibbs sampling using inversion method
x=matrix(-5,2000); y=matrix(-5,2000)
for(i in 2:2000)
 {
  #sample from x | y
  u=runif(1,min=0, max=1)
  x[i]=sqrt(u*(6*y[i-1]+8)+(1.5*y[i-1]+1)*(1.5*y[i-1]+1))
            -(1.5*y[i-1]+1)
  #sample from y | x
  u=runif(1,min=0,max=1)
  y[i]=sqrt((2*u*(4*x[i]+10))/3 +((2*x[i]+2)/3)*((2*x[i]+2)/3))
           - ((2*x[i]+2)/3)
  }
```

This program first sets the starting values for x and y equal to -5. Then, x is updated using the current value of y. Then, y is updated using the just-sampled value of x. (Notice how `x[i]` is computed using `y[i-1]`, whereas `y[i]` is sampled using `x[i]`.) Both are updated using the inversion method of sampling discussed above.

This algorithm produces samples from the marginal distributions for both x and y, but we can also treat pairs of x and y as draws from the joint density. We will discuss the conditions in which we can do this in greater depth shortly. Generally, however, of particular interest are the marginal distributions for parameters, since we are often concerned with testing hypotheses concerning one parameter, net of the other parameters in a model. Figure 4.5 shows a "trace plot" of both x and y as well as the marginal densities for both variables. The trace plot is simply a two-dimensional plot in which the x axis represents the iteration of the algorithm, and the y axis represents the simulated value of the random variable at each particular iteration. Heuristically, we can then take the trace plot, turn it on its edge (a 90 degree clockwise turn), and allow the "ink" to fall down along the y-axis and "pile-up" to produce a histogram of the marginal density. Places in the trace plot that are particularly dark represent regions of the density in which the algorithm simulated frequently; lighter areas are regions of the density that were more rarely visited by the algorithm. Thus, the "ink" will pile-up higher in areas for which the variable of interest has greater probability. Histograms of these marginal densities are shown to the right of their respective trace plots, with the theoretical marginal densities derived in Chapter 2 superimposed. Realize that these marginals are unnormalized, because the leading $1/28$ normalizing constant cancels in both the numerator and the denominator.

Notice that, although the starting values were very poor (-5 is not a valid point in either dimension of the density), the algorithm converged very rapidly to the appropriate region—[0, 2]. It generally takes a number of iterations for an MCMC algorithm to find the appropriate region—and, more theoretically, for the Markov chain produced by the algorithm to sample from the appropriate "target" distribution. Thus, we generally discard a number of early iterations before making calculations (called the "burn-in"). The marginal densities, therefore, are produced from only the last 1,500 iterations of the algorithm.

The histograms for the marginal densities show that the algorithm samples appropriately from the densities of interest. Of course, there is certainly some error—observe how the histograms tend to be a little too low or high here and there. This reflects sampling error, and such error is reduced by sampling more values (e.g., using 5,000 draws, rather than 2,000); we will return to this issue in the next chapter.

Aside from examining the marginal distributions for x and y, we can also examine the joint density. Figure 4.6 shows a two-dimensional trace plot, taken at several stages. The upper left figure shows the state of the algorithm after 5 iterations; the upper right figure shows the state after 25 iterations; the

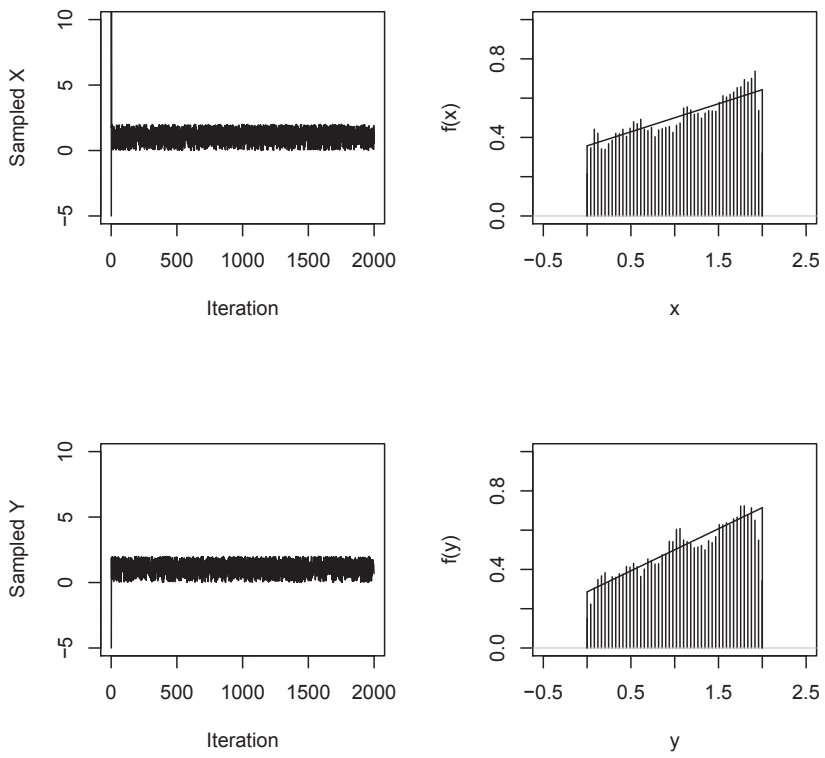

Fig. 4.5. Results of Gibbs sampler using the inversion method for sampling from conditional densities.

lower left figure shows it after 100 iterations; and the lower right figure shows it after the 2,000 iterations. Here again, we see that the algorithm, although starting with poor starting values, converged rapidly to the appropriate two-dimensional, partial plane region represented by $f(x, y)$.

After sampling from the distribution for x and y, we can now summarize our knowledge of the density. The theoretical mean for x can be found by taking the marginal for x ($f(x) = (1/28)(4x + 10)$) and by integrating across all values for x:

$$\mu_x = \int_0^2 x \times f(x)dx = 1.095.$$

A similar calculation for y yields a theoretical mean of 1.143. The empirical estimates of the means, based on the last 1,500 draws from the marginal distributions for the variables (discarding the first 500 as the burn-in) are

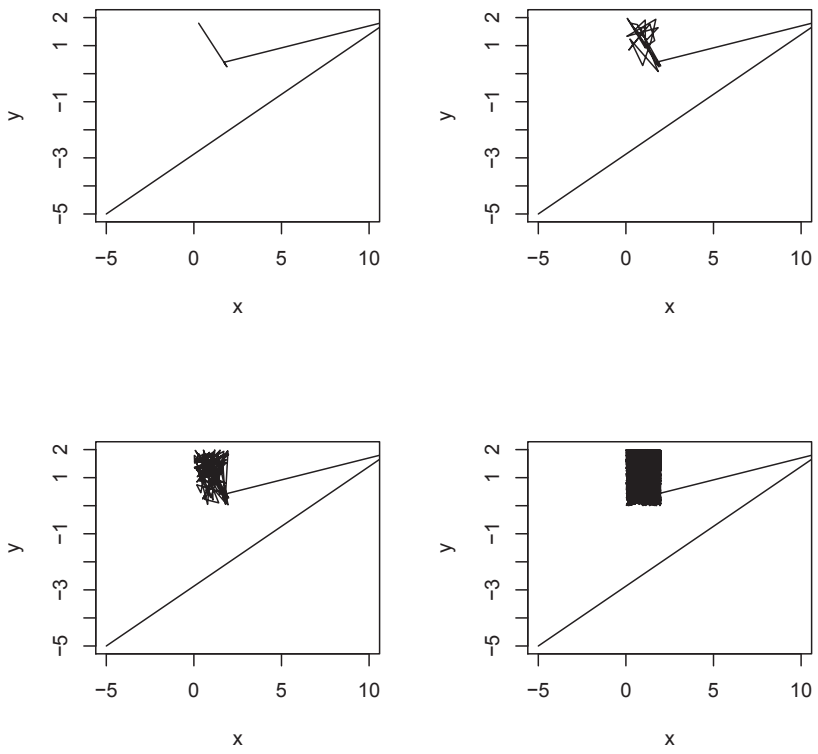

Fig. 4.6. Results of Gibbs sampler using the inversion method for sampling from conditional densities: Two-dimensional view after 5, 25, 100, and 2,000 iterations.

$\bar{x} = 1.076$ and $\bar{y} = 1.158$. The discrepancy between the theoretical and the empirical means is attributable to sampling error in the MCMC algorithm. A longer run would reduce the error, although, even with 1,500 simulated draws, the discrepancies here are minimal (less than 2% for both x and y).

4.4.3 Example repeated using rejection sampling

In the Gibbs sampling algorithm we just discussed, we used the inversion method for sampling from the conditional distributions of x and y. It is often the case that using the inversion method may not be feasible, for several reasons. First, the conditionals in the Gibbs sampler may not be univariate. That is, we do not have to break our conditional distributions into univariate conditional densities; we may choose multivariate conditional densities, as we will see in Chapter 7. Second, $F()^{-1}$ may not be calculable, even in one dimension.

For example, if the distribution were bivariate normal, the conditionals would be univariate normal, and $F()^{-1}$ cannot be analytically computed.[2] Third, even if the inverse of the density is calculable, the normalizing constant in the conditional may not be easily computable. The inversion algorithm technically requires the complete computation of $F()^{-1}$, which, in this case, requires us to know both the numerator and the denominator of the formulas for the conditional distributions. It is often the case that we do not know the exact formula for a conditional distribution, but instead, we know the conditional only up to a normalizing (proportionality) constant. Generally speaking, conditional distributions are proportional to the joint distribution evaluated at the point of conditioning. So, for example, in the example discussed above, if we know $y = q$, then the following is true:

$$f(x \mid y = q) = (1/28) \times \frac{2x + 3q + 2}{6q + 8} \propto 2x + 3q + 2.$$

Notice that $(1/28)(6q + 8)$ is not contained in the final proportionality; the reason is that this factor is simply a constant that scales this slice of the joint density so that its integral is 1. However, this constant is not necessary for Gibbs sampling to work! Why not? Because the Gibbs sampler will only set $y = q$ in direct proportion to its relative frequency in the joint density. Put another way, the Gibbs sampler will visit $y = q$ as often as it should under the joint density. This result is perhaps easier to see in a contingency table; consider the example displayed in Table 4.1.

Table 4.1. Cell counts and marginals for a hypothetical bivariate dichotomous distribution.

	$x = 0$	$x = 1$	$x \mid y = k$
$y = 0$	a	b	$a + b$
$y = 1$	c	d	$c + d$
$y \mid x = m$	$a + c$	$b + d$	$a + b + c + d$

In this example, if we follow a Gibbs sampling strategy, we would choose a starting value for x and y; suppose we chose 0 for each. If we started with $y = 0$, we would then select $x = 0$ with probability $a/(a + b)$ and $x = 1$ with probability $b/(a + b)$. Once we had chosen our x, if x had been 0, we would then select $y = 0$ with probability $a/(a + c)$ and $y = 1$ with probability $c/(a + c)$. On the other hand, if we had selected $x = 1$, we would then select

[2] Again, we *do* have efficient algorithms for computing this integral, but it cannot be directly analytically computed.

$y = 0$ with probability $b/(b + d)$ and $y = 1$ with probability $d/(b + d)$. Thus, we would be selecting $y = 0$ with total probability

$$p(y = 0) = p(y = 0 \mid x = 0)p(x = 0) + p(y = 0 \mid x = 1)p(x = 1).$$

So,

$$\begin{aligned} p(y = 0) &= \left(\frac{a}{a + c}\right)\left(\frac{a + c}{a + b + c + d}\right) + \left(\frac{b}{b + d}\right)\left(\frac{b + d}{a + b + c + d}\right) \\ &= \frac{a + b}{a + b + c + d}. \end{aligned}$$

This proportion reflects exactly how often we should choose $y = 0$, given the marginal distribution for y in the contingency table. Thus, the normalizing constant is not relevant, because the Gibbs sampler will visit each value of one variable in proportion to its relative marginal frequency, which leads us to then sample the other variable, conditional on the first, with the appropriate relative marginal frequency.

Returning to the example at hand, then, we simply need to know what the conditional distribution is proportional to in order to sample from it. Here, if we know $y = q$, then $f(x \mid y = q) \propto 2x + 3q + 2$. Because we do not necessarily always know this normalizing constant, using the inversion method of sampling will not work.[3] However, we can simulate from this density using rejection sampling. Recall from the discussion of rejection sampling that we need an enveloping function $g(x)$ that, when multiplied by a constant m, returns a value that is greater than $f(x)$ for all x. With an unnormalized density, only m must be adjusted relative to what it would be under the normalized density in order to ensure this rule is followed. In this case, if we will be sampling from the joint density, we can use a uniform density on the $[0, 2]$ interval multiplied by a constant m that ensures that the density does not exceed $m \times .5$ (.5 is the height of the U(0,2) density). The joint density reaches a maximum where x and y are both 2; that peak value is 12. Thus, if we set $m = 25$, the $U(0, 2)$ density multiplied by m will always be above the joint density. And, we can ignore the normalizing constants, including the leading $(1/28)$ in the joint density and the $1/(6y + 8)$ in the conditional for x and the $1/(4x + 10)$ in the conditional for y. As exemplified above, the Gibbs sampler will sample from the marginals in the correct proportion to their relative frequency in the joint density. Below is a Gibbs sampler that simulates from $f(x, y)$ using rejection sampling:

[3] The normalizing constant must be known one way or another. Certainly, we can perform the integration we need to compute F^{-1} so long as the distribution is proper. However, if we do not know the normalizing constant, the integral will differ from 1, which necessitates that our uniform draw representing the area under the curve be scaled by the inverse of the normalizing constant in order to represent the area under the unnormalized density fully.

```
#R program for Gibbs sampling using rejection sampling
x=matrix(-1,2000); y=matrix(-1,2000)
for(i in 2:2000)
 {
  #sample from x | y using rejection sampling
  z=0
  while(z==0)
   {
    u=runif(1,min=0, max=2)
    if( ((2*u)+(3*y[i-1])+2) > (25*runif(1,min=0,max=1)*.5))
     {x[i]=u; z=1}
   }
  #sample from y | x using rejection sampling
  z=0
  while(z==0)
   {
    u=runif(1,min=0,max=2)
    if( ((2*x[i])+(3*u)+2) > (25*runif(1,min=0,max=1)*.5))
     {y[i]=u; z=1}
   }
 }
```

In this program, the overall Gibbs sampling process is the same as for the inversion sampling approach; the only difference is that we are now using rejection sampling to sample from the unnormalized conditional distributions. One consequence of switching sampling methods is that we have now had to use better starting values (-1 here versus -5 under inversion sampling). The reason for this is that the algorithm will never get off the ground otherwise. Notice that the first item to be selected is $x[2]$. If $y[1]$ is -5, the first conditional statement (if ...) will never be true: The value on the left side of the expression, `((2*u)+(3*y[i-1])+2),` can never be positive, but the value on the right, `(25*runif(1,min=0,max=1)*.5),` will always be positive. So, the algorithm will "stick" in the first `while` loop.

Figures 4.7 and 4.8 are replications of the previous two figures produced under rejection sampling. The overall results appear the same. For example, the mean for x under the rejection sampling approach was 1.085, and the mean for y was 1.161, which are both very close to those obtained using the inversion method.

4.4.4 Gibbs sampling from a real bivariate density

The densities we examined in the examples above were very basic densities (linear and planar) and are seldom used in social science modeling. In this section, I will discuss using Gibbs sampling to sample observations from a density that *is* commonly used in social science research—the bivariate normal density. As discussed in Chapter 2, the bivariate normal density is a special case of the multivariate normal density in which the dimensionality of the

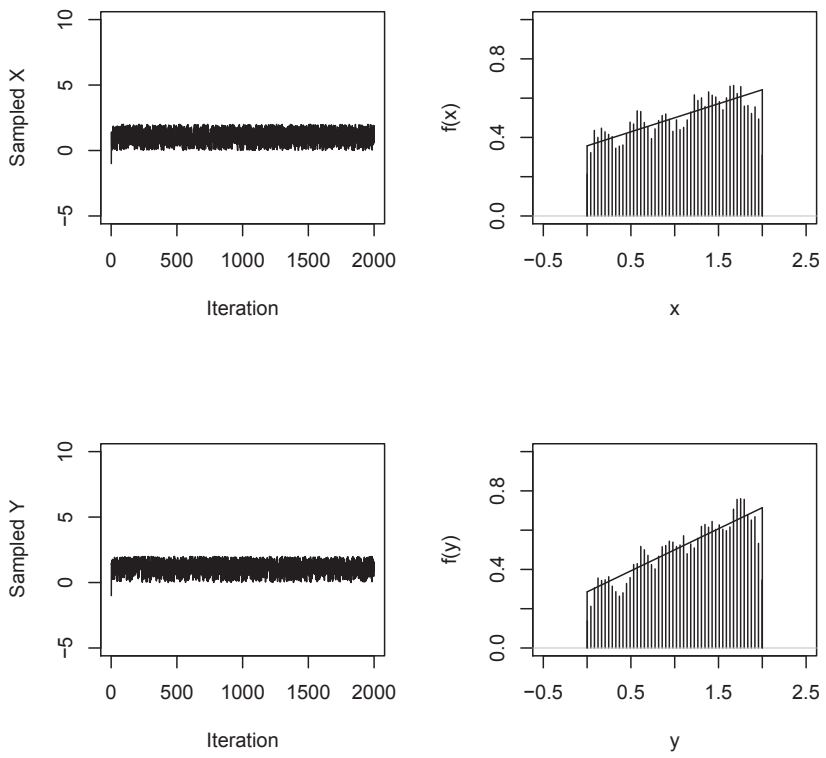

Fig. 4.7. Results of Gibbs sampler using rejection sampling to sample from conditional densities.

density is 2, and the variables—say x and y—in this density are related by the correlation parameter ρ. For the sake of this example, we will use the standard bivariate normal density—that is, the means and variances of both x and y are 0 and 1, respectively—and we will assume that ρ is a known constant (say, .5). The pdf in this case is:

$$f(x, y|\rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right\}.$$

In order to use Gibbs sampling for sampling values of x and y, we need to determine the full conditional distributions for both x and y, that is, $f(x|y)$ and $f(y|x)$. I have suppressed the conditioning on ρ in these densities, simply because ρ is a known constant in this problem.

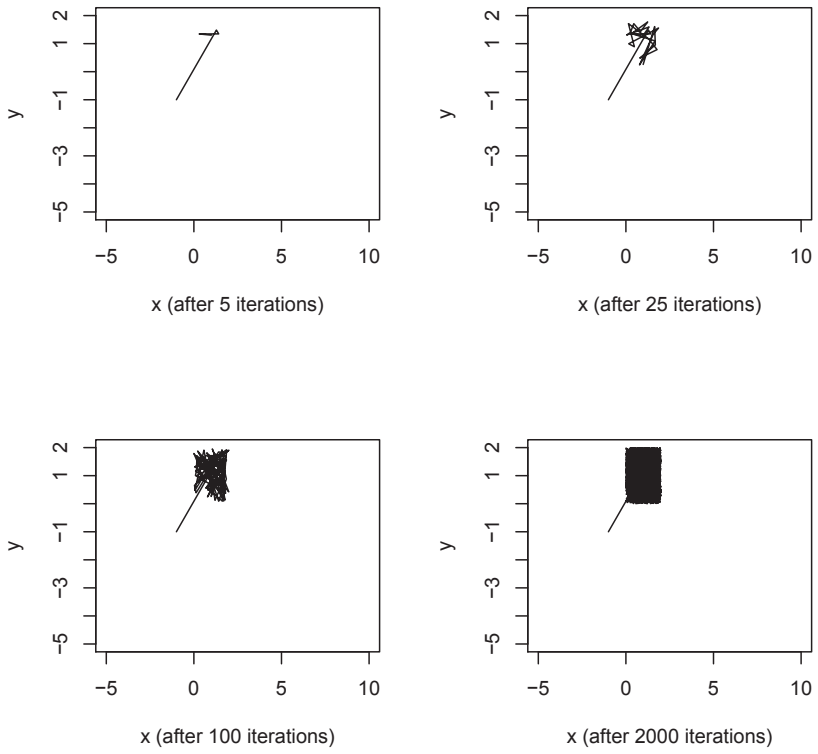

Fig. 4.8. Results of Gibbs sampler using rejection sampling to sample from conditional densities: Two-dimensional view after 5, 25, 100, and 2,000 iterations.

As we discussed above, Gibbs sampling does not require that we know the normalizing constant; we only need to know to what density each conditional density is proportional. Thus, we will drop the leading constant $(1/(2\pi\sqrt{1-\rho^2}))$. The conditional for x then requires that we treat y as known. If y is known, we can reexpress the kernel of the density as

$$f(x|y) \propto \exp\left\{-\frac{x^2 - x(2\rho y)}{2(1-\rho^2)}\right\} \exp\left\{-\frac{y^2}{2(1-\rho^2)}\right\},$$

and we can drop the latter exponential containing y^2, because it is simply a proportionality constant with respect to x. Thus, we are left with the left-hand exponential. If we complete the square in x, we obtain

$$f(x|y) \propto \exp\left\{-\frac{(x^2 - x(2\rho y) + (\rho y)^2 - (\rho y)^2)}{2(1-\rho^2)}\right\},$$

which reduces to

$$f(x|y) \propto \exp\left\{-\frac{(x-\rho y)^2-(\rho y)^2}{2(1-\rho^2)}\right\}.$$

Given that both ρ and y are constants in the conditional for x, the latter term on the right in the numerator can be extracted just as y^2 was above, and we are left with:

$$f(x|y) \propto \exp\left\{-\frac{(x-\rho y)^2}{2(1-\rho^2)}\right\}.$$

Thus, the full conditional for x can be seen as proportional to a univariate normal density with a mean of ρy and a variance of $(1-\rho^2)$. We can find the full conditional for y exactly the same way. By symmetry, the full conditional for y will be proportional to a univariate normal density with a mean of ρx and the same variance.

Writing a Gibbs sampler to sample from this bivariate density, then, is quite easy, especially given that R (and most languages) have efficient algorithms for sampling from normal distributions (`rnorm` in R). Below is an R program that does such sampling:

```
#R program for Gibbs sampling from a bivariate normal pdf
x=matrix(-10,2000); y=matrix(-10,2000)
for(j in 2:2000)
 {
  #sampling from x|y
  x[j]=rnorm(1,mean=(.5*y[j-1]),sd=sqrt(1-.5*.5))
  #sampling from y|x
  y[j]=rnorm(1,mean=(.5*x[j]),sd=sqrt(1-.5*.5))
 }
```

This algorithm is quite similar to the Gibbs sampler shown previously for the bivariate planar density. The key difference is that the conditionals are normal; thus, x and y are updated using the `rnorm` random sampling function.

Figure 4.9 shows the state of the algorithm after 10, 50, 200, and 2,000 iterations. As the figure shows, despite the poor starting values of -10 for both x and y, the algorithm rapidly converged to the appropriate region (within 10 iterations).

Figure 4.10 contains four graphs. The upper graphs show the marginal distributions for x and y for the last 1,500 iterations of the algorithm, with the appropriate "true" marginal distributions superimposed. As these graphs show, the Gibbs sampler appears to have generated samples from the appropriate marginals. In fact, the mean and standard deviation for x are .059 and .984, respectively, which are close to their true values of 0 and 1. Similarly, the mean and standard deviation for y were .012 and .979, which are also close to their true values.

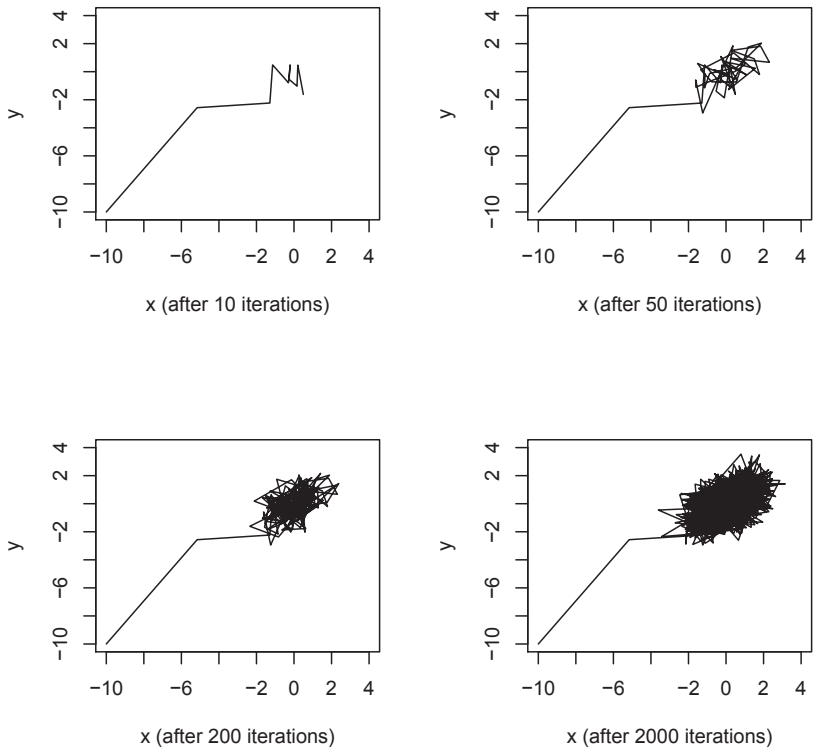

Fig. 4.9. Results of Gibbs sampler for standard bivariate normal distribution with correlation $r = .5$: Two-dimensional view after 10, 50, 200, and 2,000 iterations.

As I said earlier, we are typically interested in just the marginal distributions. However, I also stated that the samples of x and y can also be considered—after a sufficient number of burn-in iterations—as a sample from the joint density for both variables. Is this true? The lower left graph in the figure shows a contour plot for the true standard bivariate normal distribution with correlation $r = .5$. The lower right graph shows this same contour plot with the Gibbs samples superimposed. As the figure shows, the countour plot is completely covered by the Gibbs samples.

4.4.5 Reversing the process: Sampling the parameters *given* the data

Sampling *data* from densities, conditional on the parameters of the density, as we did in the previous section is an important process, but the process of Bayesian statistics is about sampling *parameters* conditional on having data,

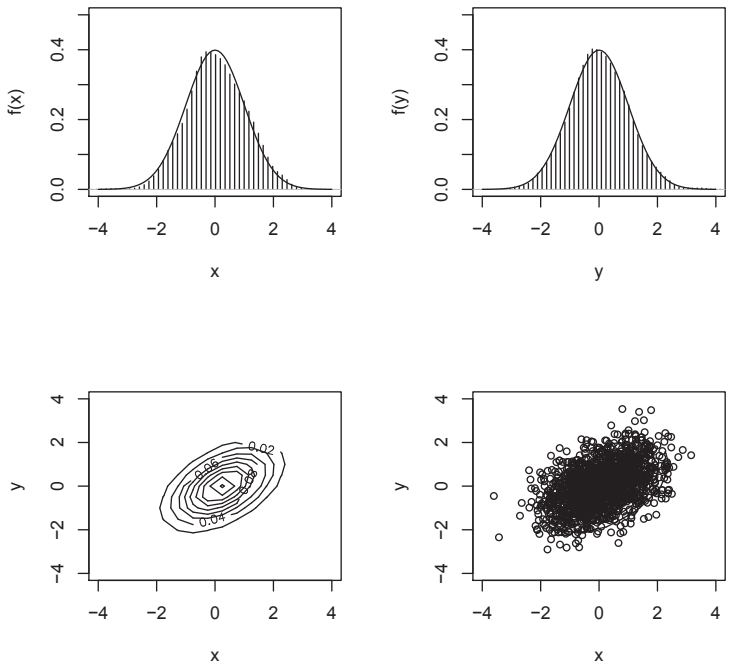

Fig. 4.10. Results of Gibbs sampler for standard bivariate normal distribution: Upper left and right graphs show marginal distributions for x and y (last 1,500 iterations); lower left graph shows contour plot of true density; and lower right graph shows contour plot of true density with Gibbs samples superimposed.

not about sampling *data* conditional on knowing the parameters. As I have repeatedly said, however, from the Bayesian perspective, both data and parameters are considered random quantities, and so sampling the parameters conditional on data is not a fundamentally different process than sampling data conditional on parameters. The main difference is simply in the mathematics we need to apply to the density to express it as a conditional density for the parameters rather than for the data. We first saw this process in the previous chapter when deriving the conditional posterior distribution for the mean parameter from a univariate normal distribution.

Let's first consider a univariate normal distribution example. In the previous chapter, we derived two results for the posterior distributions for the mean and variance parameters (assuming a reference prior of $1/\sigma^2$). In one, we showed that the posterior density could be factored to produce (1) a marginal posterior density for σ^2 that was an inverse gamma distribution, and (2) a conditional posterior density for μ that was a normal distribution:

$$p(\sigma^2|X) \propto IG\left((n-1)/2\ ,\ (n-1)\text{var}(x)/2\right)$$
$$p(\mu|\sigma^2, X) \propto N\left(\bar{x}\ ,\ \sigma^2/n\right).$$

In the second derivation for the posterior distribution for σ^2, we showed that the *conditional* (not marginal) distribution for σ^2 was also an inverse gamma distribution, but with slightly different parameters:

$$p(\sigma^2|\mu, X) \propto IG\left(n/2\ ,\ \sum(x_i-\mu)^2/2\right).$$

Both of these derivations lend themselves easily to Gibbs sampling. Under the first derivation, we could first sample a vector of values for σ^2 from the marginal distribution and then sample a value for μ conditional on each value of σ^2 from its conditional distribution. Under the second derivation, we would follow the iterative process shown in the previous sections, first sampling a value for σ^2 conditional on μ, then sampling a value for μ conditional on the new value for σ^2, and so on.

In practice, the first approach is more efficient. However, some situations may warrant the latter approach (e.g., when missing data are included). Here, I show both approaches in estimating the average years of schooling for the adult U.S. population in 2000. The data for this example are from the 2000 National Health Interview Survey (NHIS), a repeated cross-sectional survey conducted annually since 1969. The data set is relatively large by social science standards, consisting of roughly 40,000 respondents in each of many years. In 2000, after limiting the data to respondents 30 years and older and deleting observations missing on education, I obtained an analytic sample of 17,946 respondents. Mean educational attainment in the sample was 12.69 years (s.d. = 3.16 years), slightly below the mean of 12.74 from the 2000 U.S. Census.[4]

Below is an R program that first samples 2,000 values of the variance of educational attainment (σ^2) from its inverse gamma marginal distribution and then, conditional on each value for σ^2, samples μ from the appropriate normal distribution:

```
#R: sampling from marginal for variance and conditional for mean

x<-as.matrix(read.table("c:\\education.dat",header=F)[,1])
sig<-rgamma(2000,(length(x)-1)/2 , rate=((length(x)-1)*var(x)/2))
sig<-1/sig
mu<-rnorm(2000,mean=mean(x),sd=(sqrt(sig/length(x))))
```

[4] In calculating the mean from the census, I recoded the census categories for (1) under 9 years; (2) 9-12 years, no diploma; (3) high-school graduate or equivalent; (4) some college, no degree; (5) Associate degree; (6) Bachelor degree; and (4) graduate or professional degree to the midpoint for years of schooling and created a ceiling of 17 years, which is the upper limit for the NHIS.

This program is remarkably short, first reading the data into a vector X and then generating 2,000 draws from a gamma distribution with the appropriate shape and scale parameters. These draws are then inverted, because R has no direct inverse gamma distribution; thus, I make use of the fact that, if $1/x$ is gamma distributed with parameters a and b, then x is inverse gamma distributed with the same parameters. Finally, the program samples μ from its appropriate normal distribution.

Below is the R program for the alternative approach in which μ and σ are sequentially sampled from their conditional distributions:

```
#R: sampling from conditionals for both variance and mean

x<-as.matrix(read.table("c:\\education.dat",header=F)[,1])
mu=matrix(0,2000); sig=matrix(1,2000)
for(i in 2:2000)
 {
  sig[i]=rgamma(1,(length(x)/2),rate=sum((x-mu[i-1])^2)/2)
  sig[i]=1/sig[i]
  mu[i]=rnorm(1,mean=mean(x),sd=sqrt(sig[i]/length(x)))
 }
```

Under this approach, we must select starting values for μ and σ^2; here I use 0 and 1, respectively (assigned when the matrices are defined in R), which are far from their estimates based on the sample means. This approach also necessitates looping, as we saw in the planar density earlier.

Figure 4.11 shows the results of both algorithms. The first 1,000 draws have been discarded from each run, because the poor starting values in the second algorithm imply that convergence is not immediate. In contrast, under the first method, convergence is immediate; the first 1,000 are discarded simply to have comparable sample sizes. As the figure shows, the results are virtually identical for the two approaches.

Numerically, the posterior means for μ under the two approaches were both 12.69, and the posterior means for σ^2 were 10.01 and 10.00, respectively (the means for $\sqrt{\sigma^2}$ were both 3.16). These results are virtually identical to the sample estimates of these parameters, as they should be. A remaining question may be: What are the reasonable values for mean education in the population? In order to answer this question, we can construct a 95% "empirical probability interval" for μ by taking the 25^{th} and 975^{th} sorted values of μ from our Gibbs samples. For both approaches, the resulting interval is [12.64 , 12.73], which implies that the true population mean for years of schooling falls in this interval with probability .95.

4.5 Conclusions

As we have seen in the last two chapters, the Bayesian approach to inference involves simply summarizing the posterior density using basic sample statistics

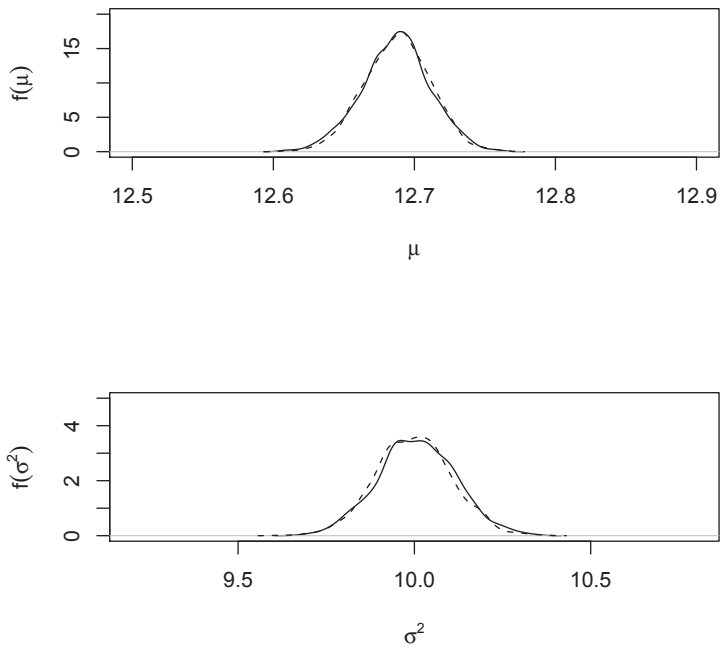

Fig. 4.11. Samples from posterior densities for a mean and variance parameter for NHIS years of schooling data under two Gibbs sampling approaches: The solid lines are the results for the marginal-for-σ^2-but conditional-for-μ approach; and the dashed lines are the results for the full conditionals approach.

like the mean, median, variance, and various quantiles of the distribution. When posterior densities are such that these integral-based statistics cannot be directly computed—e.g., when they are multivariate—modern Bayesian statistics turns to sampling from the posterior density and to computing these quantities just as we would when we have a sample of data.

Gibbs sampling provides a fairly easy method for sampling from multivariate densities, so long as we can derive the appropriate conditional densities. In most problems, this reduces simply to (1) treating other variables as fixed in the joint density, and (2) determining how to sample from the resulting conditional density. Sometimes, the conditional densities take known forms, as they did in our normal distribution example. Other times, the conditional densities may be derivable, but they may take unknown forms, as they did in our linear and planar distributions examples. In the latter case, we may turn to inversion or rejection sampling for sampling from the conditionals with unknown forms.

In some cases, however, inversion of a conditional density may not be possible, and rejection sampling may be difficult or very inefficient. In those cases, Bayesians can turn to another method—the Metropolis-Hastings algorithm. Discussion of that method is the topic of the next chapter. For alternative and more in-depth and theoretical expositions of the Gibbs sampler, I recommend the entirety of Gilks, Richardson, and Spiegelhalter 1996 in general and Gilks 1996 in particular. I also recommend a number of additional readings in the concluding chapter of this book.

4.6 Exercises

1. Find the inverse distribution function (F^{-1}) for $y|x$ in the bivariate planar density; that is, show how a $U(0,1)$ sample must be transformed to be a draw from $y|x$.
2. Develop a rejection sampler for sampling data from the bivariate planar density $f(x) \propto 2x + 3y + 2$.
3. Develop an inversion sampler for sampling data from the linear density $f(x) \propto 5x + 2$. (Hint: First, find the normalizing constant, and then find the inverse function).
4. Develop an appropriate routine for sampling the λ parameter from the Poisson distribution voting example in the previous chapter.
5. Develop an appropriate routine for sampling 20 observations (data points) from an $N(0,1)$ distribution. Then, reverse the process using these data to sample from the posterior distribution for μ and σ^2. Use the noninformative prior $p(\mu\ ,\ \sigma^2) \propto 1/\sigma^2$, and use either Gibbs sampler described in the chapter. Next, plot the posterior density for μ, and superimpose an appropriate t distribution over this density. How close is the match? Discuss.
6. As we have seen throughout this chapter, computing integrals (e.g., the mean and variance) using sampling methods yields estimates that are not exact in finite samples but that become better and better estimates as the sample size increases. Describe how we might quantify how much sampling error is involved in estimating quantities using sampling methods (Hint: Consider the Central Limit Theorem).